
Impacto de variables socioeconómicas en el desempeño estudiantil

Impact of socioeconomic variables on student performance

Manuel Arturo Jiménez Ramírez^a
manuel.jimenezr@upb.edu.co

Edward Parra^b
edward.parra@javeriana.edu.co

Mariana Luzardo Briceño^c
mariana.luzardo@upb.edu.co

Resumen

Se pretende identificar elementos explicativos a la variabilidad observada en los resultados obtenidos de las pruebas Saber Pro en los módulos de Comunicación Escrita, Inglés, Lectura Crítica, Razonamiento Cuantitativo y Competencias Ciudadanas a partir de las bases de datos suministradas por el Instituto Colombiano para el Fomento de la Educación Superior en el periodo comprendido entre 2012-10 y 2015-30 de los estudiantes de la Universidad Pontificia Bolivariana de Colombia. Se analizaron variables de tipo binario correspondientes a la existencia o no de recursos como internet, televisión, telefonía, lavadora, horno, microondas, nevera, automóvil y reproductor DVD en los hogares en los que reside el estudiante. En virtud de la gran cantidad de datos con la que se cuenta se seleccionó la técnica de Clúster Bietápico para conocer la agrupación de la información. Con éstos se realizó un paralelo con los puntajes obtenidos en la prueba Saber Pro y se encontró que, en su mayoría, los estudiantes con mayores recursos económicos obtienen los mejores puntajes en las Pruebas en cuestión.

Abstract

This document aims to identify explanatory elements to the variability observed in the results of the Saber Pro tests in the modules of Written Communication, English, Critical Reading, Quantitative Reasoning and Citizen Competences from databases supplied by the Colombian Institute for the Promotion of Higher Education Institute in the period between 2012-10 y 2015-30 of students of Pontifical Bolivarian University. It was analyzed variables of binary type corresponding to

^aUniversidad Pontificia Bolivariana, Bucaramanga

^bPontificia Universidad Javeriana

^cUniversidad Pontificia Bolivariana Bucaramanga

the existence or not of resources such as Internet, Television, Telephony, Washing Machine, Oven, Microwave, Fridge, Car, and DVD Player in the homes in which the student resides. In virtue of the large amounts of data the two-step-cluster analysis technique was selected to know the grouping of the information. With these results a parallel was made with the scores obtained in the Saber Pro test and it was found that, mostly, the students with more economic resources obtain better scores in the test.

Keywords: Cluster analysis, Classification methods, Analysis of Education, Social and Economic Stratification..

1. Introducción

La autoevaluación institucional con fines de acreditación está adaptando nuevas formas de evaluación, obligando a las instituciones a redefinir sus procesos académicos y curriculares con el fin de favorecer la equidad en la enseñanza de las diferentes disciplinas. Según Field et al. (2007) la equidad hace referencia a hacer posible que la mayoría de estudiantes desarrollen competencias y rindan bien en las instituciones, así como adquiriendo habilidades básicas y a realizarse como seres humanos, ya que dentro de los pilares se encuentra en impacto social que generan las instituciones. Field et al. (2007) arguye que equidad en educación incluye dos dimensiones, justicia e inclusión: Justicia, implica que las circunstancias personales y sociales como el género, estatus socioeconómico u origen étnico no deberían ser un obstáculo para el éxito educativo; inclusión, implica un estándar mínimo de educación para todos.

Una de las últimas metas de los diseñadores de políticas públicas es la de permitir a los ciudadanos tomar ventaja de una economía mundial globalizada; esto los está guiando a enfocarse en la mejora de las políticas educativas, asegurando la calidad en la prestación del servicio, una mayor distribución de igualdad de oportunidades de aprendizaje e iniciativas más fuertes para mayor eficiencia en las escuelas (for Economic Co-operation & (OECD) 2010).

Como lo mencionan Field et al. (2007), equidad es diferente de igualdad. Esta está asociada con ideas más amplias sobre justicia, algunas veces con “igualdad de oportunidad” y otras con “trato equivalente”. Para ellos equidad en educación incluye dos dimensiones: justicia e inclusión. Justicia, implica que las circunstancias personales y sociales como el género, estatus socioeconómico u origen étnico no deberían ser un obstáculo para el éxito educativo; inclusión, implica un estándar mínimo de educación para todos.

En este orden de ideas, el estatus socioeconómico es una medida de ventaja relativa, que explica las múltiples características de los antecedentes, incluyendo ingreso familiar, educación y ocupación de los padres.

Brown et al. (2017) en su investigación sobre el estatus socioeconómico de los estudiantes, encontró que solo 1 de cada 10 ellos en condición económica no favorecida

se ubica en el cuartil más alto de desempeño, comparado con cerca de la mitad de los estudiantes de hogares más privilegiados. Dentro del cuartil más alto de logro, hubo pequeñas diferencias significativas de desempeño (resultado promedio) entre los grupos de estudiantes de estatus socioeconómico alto y bajo. Específicamente, en pruebas de lectura y matemáticas, los estudiantes con alto desempeño y de estatus socioeconómico bajo respondieron cerca de 1.5 de preguntas menos correctamente que los estudiantes de alto desempeño con estatus socioeconómico alto, en promedio.

Malas habilidades básicas significan menos oportunidad de un trabajo, mala salud, más criminalidad y una vida más corta. La evidencia muestra que el riesgo es más agravado para aquellos menos favorecidos y que reciben una educación débil (OECD 2012).

Por supuesto, no todos los estudiantes pueden tener que rendir bien en la escuela, pero la meta de equidad en educación es la de asegurar que tantos, como sea posible, lo hagan - adquirir habilidades básicas y adicionales, realizarse como seres humanos superando accidentes de circunstancias personales y contexto familiar - no hay inevitabilidad para el fracaso en la educación. En Finlandia, y a través de otras iniciativas en muchos países diferentes, el fracaso y el abandono escolar puede ser abordado exitosamente (Field et al. 2007).

Reardon (2011) analiza la brecha entre el ingreso familiar y logro académico de los estudiantes, a partir de la consecución de los estudios educativos de los padres, concluyendo que, aunque un aumento en la desigualdad de ingresos puede jugar un rol en el crecimiento de la brecha de desempeño, éste no parece ser un factor dominante. La brecha parece haber crecido, al menos en parte, por un incremento en la asociación ingreso familiar y desempeño académico del estudiante para familias por encima de la media del nivel de ingresos.

En este sentido, las instituciones de educación superior están reconociendo cada vez más que son parte de las industrias de servicios con los estudiantes como el consumidor principal (Alfiani et al. 2015). Uno de los indicadores de desempeño académico estudiantil es el promedio acumulado de los estudiando que dinámicamente va cambiando semestre a semestre (Alfiani et al. 2015).

Mejorar el desempeño académico requiere información relacionada con las condiciones socio-demográficas del estudiante, en virtud que los estudiantes tienen diferentes niveles de motivación, entendimientos de la enseñanza y aprendizaje y, las condiciones de su entorno son diferentes entre unos y otros. Por consiguiente, las técnicas de minería de datos no son solo útiles para mapear a los estudiantes basado en variables demográficas, el entendimiento de proceso de aprendizaje, el nivel de actividad, sino también puede ser utilizada para conocer las variables que influyen en el nivel de desempeño de logro de los estudiantes en las pruebas Saber Pro como el promedio acumulado, puntajes de pruebas, duración del estudio, entre otros. El Instituto Colombiano para el Fomento de la Educación Superior (ICFES), es una entidad pública adscrito al Ministerio de Educación Nacional, que evalúa el sistema educativo colombiano en todos sus niveles y modalidades. Dentro de las pruebas que realiza se encuentran las Saber Pro, a partir de la cual

se conocen los niveles de competencias específicas y genéricas de los estudiantes universitarios a nivel nacional.

Las pruebas Saber Pro son evaluaciones censales que realiza el Ministerio de Educación para conocer los niveles de competencias de los estudiantes a nivel nacional. Se evalúan tanto competencias específicas como genéricas para los programas académicos que se ofrecen en las instituciones. Para esta investigación, y para poder establecer un comparativo entre programas se utilizaron los puntajes de los módulos genéricos que son: Comunicación Escrita, Inglés, Lectura Crítica, Razonamiento Cuantitativo y competencias ciudadanas.

En este trabajo se analiza si las diferentes variables socioeconómicas, departamento de origen, nivel educativo de los padres y el programa académico según grupo de referencia afectan el puntaje obtenido en las pruebas Saber Pro de los estudiantes de la Universidad Pontificia Bolivariana en el periodo comprendido entre 2012-10 y 2015-30, con la finalidad de tener un amplio panorama de las políticas internas que se deban formular para mejorar la calidad de los egresados.

2. Materiales y métodos

El presente es un estudio con enfoque cuantitativo, de tipo descriptivo, correlacional y diseño no experimental. Para esta investigación, y para poder establecer un comparativo entre grupos de referencia, se utilizaron los puntajes de los módulos genéricos de la prueba en cuestión, a saber, Comunicación Escrita, Inglés, Lectura Crítica, Razonamiento Cuantitativo y Competencias Ciudadanas. La base de datos utilizada fue suministrada por el ICFES, de la cual se tomaron 11.554 registros de siete semestres, entre 2012-10 y 2015-30.

Luego de analizar la base de datos se procedió a realizar un análisis *clúster* para conocer las características particulares que tenían los estudiantes en relación a los resultados en la prueba. En virtud que la base de datos tiene en su mayoría variables de tipo categórico, se utilizaron las variables que contenían información binaria con valores 1 (presencia) y 0 (ausencia) para realizar el análisis *clúster* de una manera más precisa. Las variables son las siguientes:

A partir de estas variables se procedió a utilizar el análisis de *clúster* bietápico debido a que es exploratorio y no requiere especificaciones a priori del número de clústeres (Sideridis et al. 2006), reduce el sesgo potencial si se utilizara un solo método (Ferguson et al. 2000), incrementa la validez de la solución (Punj & Stewart 1983); y asume independencia de las variables de tal forma que, la distribución conjunta multinomial-normal puede ser ubicada tanto en variables categóricas como continuas (Brown et al. 2017).

El algoritmo de *clúster* bietápico utiliza el logaritmo de máxima verosimilitud, el cual es una medida de distancia basada en la probabilidad (IBM Knowledge Center 2012). Cuando se calcula el logaritmo de la verosimilitud se asumen distribuciones multinomiales para variables categóricas, y también que las variables

Tabla 1: Variables socioeconómicas

Variable	Definición
Econ_sn_celular	El hogar cuenta con celular
Econ_sn_internet	El hogar cuenta con conexión a internet
Econ_sn_servicio_tv	El hogar cuenta con servicio cerrado de televisión
Econ_sn_telefonia	El hogar cuenta con servicio de teléfono fijo
Econ_sn_lavadora	El hogar cuenta con lavadora
Econ_sn_horno	El hogar cuenta con horno eléctrico o gas
Econ_sn_microondas	El hogar cuenta con horno microondas
Econ_sn_nevera	El hogar cuenta con nevera o enfriador
Econ_sn_automovil	El hogar cuenta con automóvil particular
Econ_sn_reproductordvd	El hogar cuenta con reproductor DVD

Fuente: Diccionario de variables, ICFES, pruebas Saber Pro.

son independientes entre sí (IBM Knowledge Center 2012). La distancia entre un clúster j y un clúster i está relacionada con la disminución en el logaritmo de verosimilitud a medida que se combinan en un solo clúster, su fórmula está definida por:

$$d(i, j) = \xi_i + \xi_j - \xi_{\langle i, j \rangle}, \quad (1)$$

donde

$$\xi_v = -N_v \left(\sum_{k=1}^{K^A} \frac{1}{2} \log(\hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2) + \sum_{k=1}^{K^B} \hat{E}_{vk} \right), \quad (2)$$

y

$$\hat{E}_{vk} = - \sum_{l=1}^{L_k} \frac{N_{vkl}}{N_v} \log \frac{N_{vkl}}{N_v}. \quad (3)$$

Si $\hat{\sigma}_k^2$ es ignorado en la expresión ξ_v , la distancia entre los *clústeres* i y j va a ser exactamente la disminución en el logaritmo de la verosimilitud cuando dos clústeres son combinados; el término $\hat{\sigma}_k^2$ es añadido para solucionar el problema causado por $\hat{\sigma}_{vk}^2 = 0$, lo que resultaría en que el logaritmo natural fuera indefinido (IBM Knowledge Center 2012).

La selección del modelo se realizó utilizando el criterio de información bayesiana que se computa de la siguiente forma:

$$BIC(J) = -2 \sum_{j=1}^J \xi_j + m_J \log(N), \quad (4)$$

donde

$$m_J = J\{2K^A + \sum_{k=1}^{K^B} (L_K - 1)\}. \quad (5)$$

El algoritmo *clúster* utiliza un procedimiento en dos etapas que funciona satisfactoriamente con el método jerárquico. En este sentido, en la primera etapa se calcula el Criterio de Información Bayesiano *BIC* para cada número de *clústeres* dentro de un rango específico, y utilizado para encontrar el estimador inicial para cada número de *clústeres* (IBM Knowledge Center 2012).

En el segundo paso, el estimador inicial es redefinido encontrando el incremento relativo más grande en distancia entre en los dos *clústeres* más cercanos en cada etapa del análisis de *clúster* jerárquico (IBM Knowledge Center 2012).

El *BIC* es uno de los criterios más populares para la selección de modelos, tal como lo menciona (Zhao et al. 2015), pero suele penalizar la complejidad de cada componente utilizando todo el tamaño de muestra; la solución a este problema se encuentra en el uso del análisis de *clúster* jerárquico en la segunda etapa, la cual tiene como clave que los parámetros son penalizados únicamente en base a sus tamaños de muestra efectivos relevantes.

Para realizar el análisis *clúster* se utilizó el software SPSS Modeler. Las variables seleccionadas (tabla 1) se procesaron siguiendo el algoritmo bietápico, con el *BIC* como criterio de selección del número de clústeres.

Una vez obtenidos los *clústeres* se realizó un análisis descriptivo de la información para cada uno de ellos en relación con: módulo de referencia, semestre en que presentó la prueba, estrato socioeconómico, departamento de origen, nivel de educación del padre, nivel de educación de la madre, género e institución para conocer el desempeño entre y al interior de cada *clúster*. La comparación se realizó utilizando como categoría de referencia el *clúster* 1.

En una segunda fase se utilizó la regresión con variables categóricas para analizar la existencia de diferencias significativas del puntaje por los módulos genéricos que evalúa la prueba. Los modelos de regresión están realizando una prueba de comparación de medias a través de la utilización de variables dummy para cada una de las categorías en cada variable. Para conocer si la diferencia en las medias es significativa, se enfocará el análisis en estadístico *t* asociado a cada parámetro en el modelo (cada variable dummy tiene un parámetro asociado).

Variable dependiente:

PS = Puntaje obtenido en las pruebas Saber Pro para cada uno de los módulos genéricos evaluados (Módulo de Comunicación Escrita, Módulo de Inglés, Módulo de Lectura Crítica, Módulo de Razonamiento Cuantitativo, Módulo de Competencias Ciudadanas).

Variables independientes:

Tabla 2: Definición de variables independientes

	Variable	Definición
C	Clúster	1,2,3
I	Institución	1710, 1723, 1727, 1730
S	Semestre	20121, 20123, 20131, 20133, 20142, 20143, 20153
E	Estrato	1, 2, 3, 4, 5, 6, 8 (vive en una zona rural donde no hay estratificación económica)
GR	Grupo de referencia	Administración de afines Arquitectura y urbanismo Bellas artes y diseño Ciencias sociales Comunicación, periodismo y publicidad Derecho Economía Educación Enfermería Humanidades Ingeniería Medicina Psicología Tecnológico en ingeniería, industria y minas
	0	Ninguno
	1	No tuvo escuela
	2	Preescolar
	3	Básica primaria
	4	Básica secundaria
	5	Media vocacional
EM/EP	6	Tecnológico o técnico
Educación de la madre/	7	Universitario
Educación del padre	8	Postgrado
	9	Primaria incompleta
	10	Primaria completa
	11	Secundaria (bachillerato) incompleta
	12	Secundaria (bachillerato) completa
	13	Educación técnica o tecnológica incompleta
	14	Educación técnica o tecnológica completa
	15	Educación profesional incompleta
	16	Educación profesional completa
	17	Postgrado
	99	No sabe

Los modelos resultantes para analizar las diferencias son los siguientes:

Modelo 1:

- Entre clústeres, utilizando como categoría de referencia el clúster 1:

$$PS_i = \beta_1 + \beta_2 C_{2i} + \beta_3 C_{3i} + e_i. \quad (6)$$

Modelo 2:

- Entre institución, utilizando como categoría de referencia la institución 1 (institución 1710, central):

$$PS_i = \beta_1 + \beta_2 I_{2i} + \beta_3 I_{3i} + \beta_4 I_{4i} + e_i. \quad (7)$$

Modelo 3:

- Entre semestre de presentación de la prueba Saber Pro, utilizando como categoría de referencia el semestre 20123 (segundo semestre del año 2012):

$$PS_i = \beta_1 + \beta_2 S_{2i} + \beta_3 S_{3i} + \beta_4 S_{4i} + \beta_5 S_{5i} + \beta_6 S_{6i} + \beta_7 S_{7i} + e_i. \quad (8)$$

Modelo 4:

- Para estrato socioeconómico según factura de energía, utilizando como categoría de referencia el estrato 6:

$$PS_i = \beta_1 + \beta_2 E_{2i} + \beta_3 E_{3i} + \beta_4 E_{4i} + \beta_5 E_{5i} + \beta_6 E_{6i} + \beta_7 E_{7i} + e_i. \quad (9)$$

Modelo 5:

- Para grupo de referencia de programas académicos, utilizando como categoría de referencia Ingeniería:

$$PS_i = \beta_1 + \beta_2 GR_{2i} + \beta_3 GR_{3i} + \beta_4 GR_{4i} + \beta_5 GR_{5i} + \beta_6 GR_{6i} + \beta_7 G_{7i} + \beta_8 G_{8i} + \beta_9 G_{9i} \\ + \beta_{10} G_{10i} + \beta_{11} G_{11i} + \beta_{12} G_{12i} + \beta_{13} G_{13i} + \beta_{14} G_{14i} + e_i. \quad (10)$$

3. Resultados

Se presenta en primer lugar, el resultado del número de clústeres seleccionados, luego el análisis descriptivo de las variables independientes en cada uno de ellos, y, por último, los modelos de regresión que permiten conocer las diferencias significativas entre los puntajes promedio según el caso, a saber, clúster, sede de la institución, semestre de presentación de la prueba, estrato socioeconómico y grupo de referencia.

3.1. Número de clústeres seleccionados

Tal como lo muestra la salida del software, con las nueve variables de entrada que se observan en la tabla 1, el usar tres clústeres la calidad de los mismos es correcta.

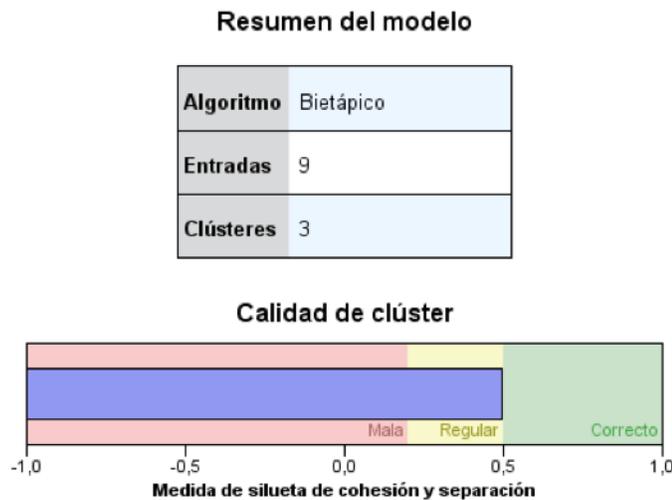


Figura 1: Resultados y selección de número de clústeres.
 Fuente: Elaboración propia utilizando SPSS Modeler, 2017

3.2. Análisis descriptivo de las variables independientes por clúster

El análisis descriptivo se realizó con apoyo del software SPSS Modeler, caracterizando variables relacionadas con el estudiante para cada uno de los clústeres seleccionados. Los clústeres obtenidos tienen tamaños diferentes, según la presencia o ausencia de diferentes recursos medidos en la base de datos. El clúster 1 tiene un tamaño de 4.500 registros, el clúster 2 un tamaño de 4.905, y el clúster 3 un tamaño de 2.113.

Tabla 3: Clasificación de estudiantes por clúster según variables socioeconómicas

Variables	Clúster 1	Clúster 2	Clúster 3
Econ.sn.internet	4500	4905	961
Econ.sn.servicio.tv	4500	4905	1010
Econ.sn.telefonia	4500	4380	963
Econ.sn.lavadora	4500	4905	1076
Econ.sn.horno	4500	3623	782
Econ.sn.microondas	4500	3037	544
Econ.sn.nevera	4500	4905	1794
Econ.sn.automovil	4500	1765	446
Econ.sn.reproductordvd	4500	3278	896
Número de registros	4500	4905	2113

Fuente: Los autores.

Una vez obtenidos los registros se analizó los promedios para cada módulo evaluado en las pruebas Saber Pro, Tabla 4. Se puede apreciar que el clúster 1 tiene promedios más altos, y en todos los módulos, en relación a los clústeres 2 y 3.

Tabla 4: Promedio por módulos evaluados en las pruebas Saber 11

Módulo	Resultado en los Módulos		
	Clúster 1	Clúster 2	Clúster 3
Comunicación Escrita	10,31565119	10,23246214	9,95260116
Inglés	11,39577199	10,88533142	10,0247296
Lectura Crítica	10,58229354	10,46536015	9,96612903
Razonamiento Cuantitativo	10,46422801	10,32156782	9,76110057
Competencias Ciudadanas	10,30718133	10,16449826	9,71347249

Fuente: Los autores.

En la Tabla 5 se identifica que la Universidad ha tenido un crecimiento en su población de estudiantes en los últimos dos semestres dentro del intervalo de tiempo analizado.

Tabla 5: Número de estudiantes por semestre de presentación de la prueba Saber Pro

Semestre	Clúster 1	%	Clúster 2	%	Clúster 3	%
20121	483	10,73	455	9,28	143	6,77
20123	640	14,22	578	11,78	189	8,94
20131	416	9,24	482	9,83	240	11,36
20133	846	18,80	977	19,92	361	17,08
20142	49	1,09	51	1,04	14	0,66
20143	1029	22,87	1125	22,94	552	26,12
20153	1037	23,04	1237	25,22	614	29,06

Fuente: Los autores.

En relación al estrato socioeconómico se aprecia que la distribución de estrato al que pertenece cada estudiante cambia según el clúster; los estudiantes del clúster 1 se caracterizan por todos tener presencia de los recursos relacionados con las variables de entrada del clúster, y a su vez, la mayor proporción de estudiantes son los pertenecientes a estratos 4 y 5; en el clúster 2 se identifica que los estudiantes cuentan con la mayoría de recursos, y son pertenecientes a estratos 3 y 4; y en el clúster 3 se ubican los estudiantes que no cuentan con todos los recursos analizados, y además, en su mayoría pertenecen a estratos 1 y 3.

Tabla 6: Número de estudiantes por estrato socioeconómico

Estrato	Clúster 1	%	Clúster 2	%	Clúster 3	%
1	18	0,40	116	2,36	568	26,88
2	139	3,09	605	12,33	476	22,53
3	730	16,22	1605	32,72	538	25,46
4	1395	31,00	1596	32,54	279	13,20
5	1468	32,62	802	16,35	132	6,25
6	737	16,38	171	3,49	28	1,33
8	9	0,20	6	0,12	64	3,03

Fuente: Los autores.

En relación al departamento de procedencia la frecuencia muestra que la presencia de estudiantes de diferentes orígenes difiere según el clúster; es decir, y como se observa en la Tabla 7, el departamento de Bolívar y Boyacá no tienen presencia en el clúster 1, Boyacá no tiene presencia en el clúster 2, y el clúster 3 estos dos departamentos sí tienen presencia en este grupo. Hasta este punto se percibe que la distribución de ingresos y los recursos con los que cuenta los estudiantes afecta su desempeño académico, relación que puede apreciarse también, según DANE (2017), por la presencia de departamentos en condiciones de mayor pobreza, en el clúster 3.

La educación de los padres es otra variable que se tuvo en cuenta para conocer

Tabla 7: Número de estudiantes por departamento de origen

Departamento	Clúster 1	%	Clúster 2	%	Clúster 3	%
Amazonas	0	0,00	0	0,00	6	0,28
Antioquia	3117	76,87	2550	51,99	699	33,08
Arauca	0	0,00	0	0,00	2	0,09
Atlántico	1	0,02	2	0,04	0	0,00
Bogotá	33	0,81	63	1,28	44	2,08
Bolívar	0	0,00	1	0,02	4	0,19
Boyacá	0	0,00	0	0,00	3	0,14
Caldas	17	0,42	30	0,61	17	0,80
Caquetá	0	0,00	1	0,02	24	1,14
Casanare	0	0,00	0	0,00	2	0,09
Cauca	2	0,05	6	0,12	186	8,80
Cesar	3	0,07	2	0,04	2	0,09
Chocó	1	0,02	0	0,00	31	1,47
Córdoba	299	7,37	521	10,62	283	13,39
Cundinamarca	0	0,00	4	0,08	14	0,66
Guainía	0	0,00	2	0,04	20	0,95
Guaviare	0	0,00	0	0,00	2	0,09
Huila	1	0,02	1	0,02	3	0,14
La guajira	4	0,10	5	0,10	3	0,14
Magdalena	1	0,02	1	0,02	2	0,09
Meta	0	0,00	2	0,04	2	0,09
Nariño	0	0,00	3	0,06	25	1,18
Norte Santander	2	0,05	1	0,02	3	0,14
Putumayo	3	0,07	5	0,10	152	7,19
Quindío	1	0,02	0	0,00	0	0,00
Risaralda	2	0,05	4	0,08	3	0,14
San Andrés	1	0,02	0	0,00	0	0,00
Santander	955	23,55	1568	31,97	474	22,43
Sucre	4	0,10	8	0,16	6	0,28
Tolima	0	0,00	1	0,02	4	0,19
Valle	50	1,23	120	2,45	80	3,79
Vaupés	0	0,00	0	0,00	5	0,24
Vichada	0	0,00	0	0,00	10	0,47
(en blanco)	3	0,07	4	0,08	2	0,09

Fuente: Los autores.

cómo ésta afecta el desempeño del estudiante, en Tabla 8 (educación madre) y Tabla 9 (educación padre). En términos de frecuencia se identificó, y aunque hay porcentajes variados, los estudiantes del clúster 1 tienen padres con educación profesional completa y postgrado; los estudiantes del clúster 2 tienen padres con educación profesional completa y secundaria completa; y los estudiantes del clúster 3 tienen padres con educación primaria incompleta, seguido de estudiantes con padres con educación secundaria completa.

Tabla 8: Número de estudiantes según nivel educativo del padre

Educación Padre	Clúster 1	%	Clúster 2	%	Clúster 3	%
0	34	0,76	93	1,90	138	6,53
1	0	0,00	0	0,00	9	0,43
2	0	0,00	0	0,00	21	0,99
3	1	0,02	0	0,00	6	0,28
4	0	0,00	0	0,00	2	0,09
5	0	0,00	0	0,00	4	0,19
6	0	0,00	0	0,00	0	0,00
7	0	0,00	0	0,00	0	0,00
8	0	0,00	0	0,00	0	0,00
9	138	3,07	331	6,75	465	22,01
10	119	2,64	319	6,50	247	11,69
11	297	6,60	474	9,66	188	8,90
12	748	16,62	1059	21,59	395	18,69
13	114	2,53	134	2,73	51	2,41
14	488	10,84	602	12,27	130	6,15
15	197	4,38	212	4,32	50	2,37
16	1487	33,04	1159	23,63	295	13,96
17	877	19,49	522	10,64	112	5,30
99	0	0,00	0	0,00	0	0,00

Fuente: Los autores.

Se aprecia en la Tabla 10 que en los tres cluster la presencia femenina es superior a la masculina en al menos 7.2

En relación a las instituciones, se puede apreciar que la que tiene mayor número de estudiantes es la 1710, Tabla 11, ya que del total de la base de datos el 62.98 % de los registros le pertenecen; por esta razón es la que mayor presencia tiene en cada clúster seleccionado; en función a esto se analizó el porcentaje que corresponde según el número de estudiantes por institución y no para cada clúster, lo cual resultó en que de mayor presencia en el clúster 1 son las instituciones 1710 y 1023, en el clúster 2 se encuentran las 1723 y 1730, y en el clúster 3 las 1727 y 1730.

Con respecto al grupo de referencia se identifica que los programas de Ingeniería tienen un mayor porcentaje de estudiantes en los clústeres 1 y 2, y los programas

Tabla 9: Número de estudiantes según nivel educativo de la madre

Educación Madre	Clúster 1	%	Clúster 2	%	Clúster 3	%
0	11	0,24	37	0,75	111	5,25
1	0	0,00	0	0,00	12	0,57
2	0	0,00	0	0,00	21	0,99
3	1	0,02	0	0,00	6	0,28
4	0	0,00	0	0,00	1	0,05
5	0	0,00	0	0,00	1	0,05
6	0	0,00	0	0,00	1	0,05
7	0	0,00	0	0,00	0	0,00
8	0	0,00	0	0,00	0	0,00
9	65	1,44	212	4,32	389	18,41
10	84	1,87	223	4,55	248	11,74
11	282	6,27	435	8,87	193	9,13
12	923	20,51	1229	25,06	425	20,11
13	138	3,07	175	3,57	50	2,37
14	675	15,00	707	14,41	185	8,76
15	214	4,76	185	3,77	37	1,75
16	1375	30,56	1132	23,08	300	14,20
17	719	15,98	553	11,27	123	5,82
99	13	0,29	17	0,35	10	0,47

Fuente: Los autores.

Tabla 10: Número de estudiantes por género

Género	Clúster 1	%	Clúster 2	%	Clúster 3	%
M	1958	43,51	2101	42,83	978	46,28
F	2539	56,42	2800	57,08	1130	53,48
(En blanco)	3	0,07	4	0,08	5	0,24

Fuente: Los autores.

Tabla 11: Número de estudiantes en relación al total clúster e institución

Clúster	Institución	Número de estudiantes	Relación por total clúster	Relación por total Institución
1	1710	3207	71,27 %	44,07 %
	1723	967	21,49 %	31,77 %
	1727	300	6,67 %	26,69 %
	1730	26	0,58 %	23,85 %
2	1710	2723	55,51 %	37,42 %
	1723	1587	32,35 %	52,14 %
	1727	532	10,85 %	47,33 %
	1730	63	1,28 %	57,80 %
3	1710	1321	62,52 %	18,15 %
	1723	480	22,72 %	15,77 %
	1727	292	13,82 %	25,98 %
	1730	20	0,95 %	18,35 %
(en blanco)	1710	26	72,22 %	0,36 %
	1723	10	27,78 %	0,33 %

Fuente: Los autores.

relacionados con Educación tienen mayor presencia en el clúster 3.

Tabla 12: Número de estudiantes por grupo de referencia

Grupo de referencia	Clúster 1	%	Clúster 2	%	Clúster 3	%
ADMINISTRACIÓN Y AFINES	605	13,44	591	12,05	161	7,62
ARQUITECTURA Y URBANISMO	373	8,29	200	4,08	55	2,60
BELLAS ARTES Y DISEÑO	615	13,67	385	7,85	76	3,60
CIENCIAS SOCIALES	74	1,64	82	1,67	23	1,09
COMUNICACIÓN, PERIODISMO Y PUBLICIDAD	580	12,89	513	10,46	134	6,34
DERECHO	417	9,27	345	7,03	121	5,73
ECONOMÍA	48	1,07	83	1,69	30	1,42
EDUCACIÓN	73	1,62	307	6,26	744	35,21
ENFERMERÍA	106	2,36	207	4,22	27	1,28
HUMANIDADES	56	1,24	99	2,02	105	4,97
INGENIERÍA	1126	25,02	1606	32,74	514	24,33
MEDICINA	210	4,67	141	2,87	21	0,99
PSICOLOGÍA	216	4,80	345	7,03	101	4,78
TECNOLOGICO EN INGENIERÍA, INDUSTRIA Y MINAS	1	0,02	1	0,02	1	0,05

Al aplicar los modelos de regresión descritos en la metodología, para conocer si existe o no diferencias significativas para cada módulo entre clústeres, entre institución, entre semestre de presentación, entre, estrato y, entre grupo de referencia, se obtuvo:

Modelo 1: para clúster

Tabla 13: Parámetros de regresión por módulo prueba Saber Pro

Variable dependiente	Modelo	Coeficientes no estandarizados		Coeficientes estandarizados Beta	t	Sig.
		B	Error estándar			
Módulo	(Constante)	10,316	0,015	–	667,261	0
Comunicación	2	-0,083	0,021	-0,04	-3,888	0
Escrita	3	-0,363	0,027	-0,136	-13,281	0
Módulo	(Constante)	11,396	0,027	$-\hat{\beta}_{i, \frac{1}{2}}$	426,95	0
Inglés	2	-0,51	0,037	-0,137	-13,822	0
	3	-1,371	0,047	-0,288	-29,11	0
Módulo	(Constante)	10,582	0,018	–	580,53	0
Lectura crítica	2	-0,117	0,025	-0,047	-4,636	0
	3	-0,616	0,032	-0,193	-19,155	0
	(Constante)	10,464	0,019	$\hat{\beta}_{i, \frac{1}{2}}$	545,035	0
Módulo	2	-0,143	0,027	-0,054	-5,37	0
Razonamiento Cuantitativo	3	-0,703	0,034	-0,209	-20,754	0
Módulo	(Constante)	10,307	0,018	$-\hat{\beta}_{i, \frac{1}{2}}$	560,019	0
Competencias	2	-0,143	0,025	-0,057	-5,603	0
Ciudadanas	3	-0,594	0,032	-0,185	-18,28	0

Modelo 2: para institución

Tabla 14: Parámetros de regresión por módulo prueba Saber Pro

Variable dependiente:	Modelo	Coeficientes no estandarizados		Coeficientes estandarizados Beta	t	Sig.
		B	Error estándar			
Módulo Comunicación Escrita	(Constante)	10,268	0,012	–	838,427	0,000
	1723	-0,075	0,022	-0,032	-3,369	0,001
	1727	-0,340	0,033	-0,097	-10,196	0,000
	1730	-0,134	0,100	-0,013	-1,343	0,179
Módulo Inglés	(Constante)	11,051	0,022	–	511,910	0,000
	1723	-0,120	0,040	-0,029	-3,026	0,002
	1727	-0,799	0,059	-0,129	-13,612	0,000
	1730	-1,011	0,177	-0,053	-5,721	0,000
Módulo Lectura Crítica	(Constante)	10,451	0,014	–	720,953	0,000
	1723	0,045	0,027	0,016	1,687	0,092
	1727	-0,406	0,039	-0,098	-10,296	0,000
	1730	-0,158	0,119	-0,012	-1,330	0,183
Módulo Razonamiento Cuantitativo	(Constante)	10,197	0,015	–	671,760	0,000
	1723	0,402	0,028	0,136	14,443	0,000
	1727	-0,238	0,041	-0,054	-5,770	0,000
	1730	-0,553	0,124	-0,041	-4,449	0,000
Módulo Competencias Ciudadanas	(Constante)	10,165	0,015	–	693,193	0,000
	1723	0,018	0,027	0,007	0,686	0,493
	1727	-0,276	0,040	-0,066	-6,923	0,000
	1730	-0,290	0,120	-0,023	-2,418	0,016

En relación a la institución, se puede observar que la que tuvo mejor desempeño es la 1710 en la mayoría de los módulos, con excepción de razonamiento cuantitativo y competencias ciudadanas en la que la 1723 sobresale con resultados superiores a los de la comparación.

Modelo 3: para semestre de presentación de la prueba Saber Pro

Analizando el semestre de presentación se aprecia que el semestre en el que los estudiantes presentaron mejores desempeños en los módulos de la prueba Saber Pro fue en el periodo 20123, aunque, como se observa en la Tabla 15, unos semestres parecen mostrar que no hay diferencia significativa en relación al semestre de comparación.

Tabla 15: Parámetros de regresión por módulo prueba Saber Pro

Variable dependiente:	Modelo	Coeficientes no estandarizados		Coeficientes estandarizados Beta	t	Sig.
		B	Error estándar			
Módulo Comunicación Escrita	(constante)	10,531	0,028	–	370,203	0,000
	20121	-0,029	0,042	-0,008	-0,682	0,495
	20131	-0,530	0,042	-0,154	-12,725	0,000
	20133	-0,334	0,036	-0,127	-9,304	0,000
	20142	-0,612	0,101	-0,058	-6,053	0,000
	20143	-0,407	0,035	-0,168	-11,768	0,000
	20153	-0,375	0,034	-0,157	-10,942	0,000
Módulo Inglés	(Constante)	11,152	0,051	–	220,330	0,000
	20121	-0,178	0,076	-0,028	-2,357	0,018
	20131	-0,484	0,074	-0,078	-6,496	0,000
	20133	-0,197	0,064	-0,042	-3,065	0,002
	20142	0,448	0,161	0,027	2,785	0,005
	20143	-0,240	0,062	-0,055	-3,890	0,000
	20153	-0,269	0,061	-0,063	-4,393	0,000
Módulo Lectura Crítica	(Constante)	10,435	0,034	–	308,671	0,000
	20121	0,136	0,050	0,032	2,689	0,007
	20131	-0,197	0,050	-0,048	-3,960	0,000
	20133	-0,049	0,043	-0,016	-1,152	0,249
	20142	0,110	0,107	0,010	1,024	0,306
	20143	-0,122	0,041	-0,042	-2,962	0,003
	20153	0,120	0,041	0,042	2,927	0,003
Módulo Razonamiento Cuantitativo	(Constante)	10,374	0,036	–	290,038	0,000
	20121	-0,043	0,053	-0,010	-0,801	0,423
	20131	-0,300	0,053	-0,069	-5,691	0,000
	20133	-0,081	0,045	-0,024	-1,780	0,075
	20142	0,291	0,114	0,025	2,566	0,010
	20143	-0,120	0,044	-0,039	-2,751	0,006
	20153	-0,099	0,043	-0,033	-2,284	0,022
Módulo Competencias Ciudadanas	(Constante)	10,348	0,034	–	304,317	0,000
	20121	0,098	0,051	0,023	1,932	0,053
	20131	-0,379	0,050	-0,091	-7,565	0,000
	20133	-0,266	0,043	-0,084	-6,162	0,000
	20142	-0,006	0,108	-0,001	-0,056	0,955
	20143	-0,340	0,042	-0,116	-8,201	0,000
	20153	-0,197	0,041	-0,068	-4,789	0,000

Modelo 4: para estrato socioeconómico según factura de energía eléctrica

Con respecto al estrato socioeconómico, Tabla 16, se aprecia que los estudiantes con mejor desempeño son los estudiantes con nivel 6 de estrato socioeconómico, aunque en algunos casos se observa que para el nivel 5 la diferencia no es significativa.

Tabla 16: Parámetros de regresión por módulo prueba Saber Pro

Variable dependiente:	Modelo	Coeficientes no estandarizados		Coeficientes estandarizados Beta	t	Sig.
		B	Error estándar			
Módulo Comunicación Escrita	(Constante)	10,327	0,033	–	308,476	0,000
	1	-0,746	0,051	-0,173	-14,557	0,000
	2	-0,193	0,045	-0,058	-4,339	0,000
	3	-0,053	0,039	-0,022	-1,366	0,172
	4	-0,071	0,038	-0,031	-1,862	0,063
	5	-0,023	0,040	-0,009	-0,570	0,568
	6	-0,878	0,122	-0,069	-7,187	0,000
Módulo Inglés	(Constante)	12,054	0,057	–	212,093	0,000
	1	-2,668	0,087	-0,348	-30,808	0,000
	2	-1,848	0,075	-0,309	-24,489	0,000
	3	-1,415	0,065	-0,333	-21,644	0,000
	4	-0,938	0,064	-0,230	-14,561	0,000
	5	-0,593	0,067	-0,130	-8,850	0,000
	6	-2,498	0,204	-0,112	-12,239	0,000
Módulo Lectura Crítica	(Constante)	10,691	0,039	–	271,695	0,000
	1	-1,285	0,060	-0,250	-21,438	0,000
	2	-0,477	0,052	-0,119	-9,130	0,000
	3	-0,274	0,045	-0,096	-6,062	0,000
	4	-0,143	0,045	-0,052	-3,216	0,001
	5	-0,101	0,046	-0,033	-2,181	0,029
	6	-1,170	0,141	-0,078	-8,282	0,000
Módulo Razonamiento Cuantitativo	(Constante)	10,536	0,042	–	253,231	0,000
	1	-1,278	0,063	-0,235	-20,151	0,000
	2	-0,534	0,055	-0,126	-9,660	0,000
	3	-0,303	0,048	-0,101	-6,333	0,000
	4	-0,082	0,047	-0,028	-1,739	0,082
	5	-0,078	0,049	-0,024	-1,597	0,110
	6	-1,198	0,149	-0,076	-8,014	0,000
Módulo Competencias Ciudadanas	(Constante)	10,390	0,040	–	260,393	0,000
	1	-1,171	0,061	-0,226	-19,268	0,000
	2	-0,437	0,053	-0,108	-8,245	0,000
	3	-0,244	0,046	-0,085	-5,307	0,000
	4	-0,165	0,045	-0,060	-3,640	0,000
	5	-0,075	0,047	-0,025	-1,598	0,110
	6	-1,011	0,143	-0,067	-7,058	0,000

Modelo 5: para grupo de referencia de programas académicos

Por último, en cuanto a grupo de referencia las significancias cambian según el módulo analizado. El programa de ingeniería muestra ser más fuerte en los módulos de razonamiento cuantitativo y competencias ciudadanas, y medicina presenta mejores resultados en los módulos de lectura crítica, inglés y comunicación escrita.

Tabla 17: Parámetros de regresión por módulo prueba Saber Pro

Variable dependiente:	Modelo	Coeficientes no estandarizados		Coeficientes estandarizados Beta	t	Sig.
		B	Error estándar			
Módulo Comunicación Escrita	(Constante)	10,107	0,018	–	567,623	0,000
	Administración	0,058	0,033	0,018	1,752	0,080
	Arquitectura	0,098	0,044	0,022	2,231	0,026
	Bellas artes	0,088	0,036	0,025	2,477	0,013
	Ciencias sociales	0,508	0,078	0,061	6,515	0,000
	Comunicación	0,310	0,034	0,092	9,067	0,000
	Derecho	0,375	0,039	0,096	9,690	0,000
	Economía	0,089	0,082	0,010	1,084	0,279
	Educación	-0,398	0,035	-0,114	-11,308	0,000
	Enfermería	0,245	0,057	0,040	4,256	0,000
	Humanidades	0,712	0,065	0,103	10,914	0,000
	Medicina	0,574	0,055	0,099	10,384	0,000
	Psicología	0,220	0,043	0,049	5,059	0,000
Tecnología	1,127	0,582	0,018	1,936	0,053	
Módulo Inglés	(Constante)	11,103	0,031	–	356,511	0,000
	Administración	-0,045	0,057	-0,008	-0,775	0,438
	Arquitectura	0,252	0,077	0,031	3,285	0,001
	Bellas artes	0,330	0,062	0,052	5,303	0,000
	Ciencias sociales	-0,400	0,137	-0,027	-2,927	0,003
	Comunicación	-0,051	0,060	-0,008	-0,849	0,396
	Derecho	-0,213	0,067	-0,031	-3,161	0,002
	Economía	-0,393	0,143	-0,025	-2,758	0,006
	Educación	-1,397	0,061	-0,225	-22,768	0,000
	Enfermería	-0,943	0,101	-0,087	-9,350	0,000
	Humanidades	-0,909	0,114	-0,073	-7,965	0,000
	Medicina	1,084	0,097	0,104	11,193	0,000
	Psicología	-0,510	0,076	-0,064	-6,717	0,000
Tecnología	0,330	1,023	0,003	0,322	0,747	
Módulo Lectura Crítica	(Constante)	10,460	0,021	–	500,535	0,000
	Administración	-0,173	0,039	-0,045	-4,483	0,000
	Arquitectura	0,016	0,052	0,003	0,307	0,759
	Bellas artes	0,003	0,042	0,001	0,062	0,951
	Ciencias sociales	0,154	0,092	0,015	1,683	0,092
	Comunicación	0,200	0,040	0,050	4,993	0,000
	Derecho	0,271	0,045	0,058	5,992	0,000
	Economía	-0,118	0,096	-0,011	-1,237	0,216
	Educación	-0,910	0,041	-0,219	-22,109	0,000
	Enfermería	-0,206	0,068	-0,028	-3,050	0,002
	Humanidades	-0,023	0,077	-0,003	-0,300	0,764
	Medicina	0,819	0,065	0,117	12,597	0,000
	Psicología	0,117	0,051	0,022	2,300	0,021
Tecnología	-0,260	0,686	-0,003	-0,379	0,705	

Tabla 18: Parámetros de regresión por módulo prueba Saber Pro

Variable dependiente:	Modelo	Coeficientes no estandarizados		Coeficientes estandarizados Beta	t	Sig.
		B	Error estándar			
Módulo Razonamiento Cuantitativo	(Constante)	10,940	0,021	—	525,520	0,000
	Administración	-0,607	0,038	-0,150	-15,821	0,000
	Arquitectura	-0,527	0,051	-0,092	-10,273	0,000
	Bellas artes	-0,821	0,042	-0,183	-19,711	0,000
	Ciencias sociales	-1,125	0,091	-0,106	-12,308	0,000
	Comunicación	-1,002	0,040	-0,236	-25,105	0,000
	Derecho	-0,832	0,045	-0,169	-18,455	0,000
	Economía	-0,724	0,095	-0,065	-7,596	0,000
	Educación	-1,777	0,041	-0,405	-43,332	0,000
	Enfermería	-0,940	0,067	-0,122	-13,951	0,000
	Humanidades	-1,145	0,076	-0,131	-15,005	0,000
	Medicina	0,057	0,065	0,008	0,887	0,375
	Psicología	-1,125	0,051	-0,200	-22,177	0,000
Tecnología	-1,640	0,684	-0,020	-2,399	0,016	
Módulo Competencias Ciudadanas	(Constante)	10,177	0,021	—	483,941	0,000
	Administración	-0,144	0,039	-0,037	-3,715	0,000
	Arquitectura	-0,037	0,052	-0,007	-0,722	0,470
	Bellas artes	-0,150	0,042	-0,035	-3,577	0,000
	Ciencias sociales	0,396	0,092	0,039	4,290	0,000
	Comunicación	0,130	0,040	0,032	3,215	0,001
	Derecho	0,536	0,046	0,114	11,765	0,000
	Economía	-0,079	0,096	-0,008	-0,825	0,410
	Educación	-0,881	0,041	-0,210	-21,274	0,000
	Enfermería	-0,140	0,068	-0,019	-2,060	0,039
	Humanidades	0,053	0,077	0,006	0,684	0,494
	Medicina	0,792	0,065	0,113	12,108	0,000
	Psicología	-0,004	0,051	-0,001	-0,083	0,934
Tecnología	0,023	0,691	0,000	0,033	0,973	

4. Conclusiones

Al analizar cada uno de los clústeres se identifica diferencias significativas en los puntajes para cada uno de los módulos, con justificación en el nivel de recursos que posee el estudiante. Se puede concluir que el nivel que posee el estudiante tiene una alta relación con el estrato socioeconómico, y así mismo con la educación de los padres, y en algunos casos con el departamento de procedencia; siendo los estudiantes con mejores recursos lo que presentan mejores resultados en los módulos de la prueba Saber Pro.

La institución 1730 del clúster 3 mostraba una diferencia negativa con respecto a la 1710 del clúster 1, pero con una significancia alta lo que indicaba que esta diferencia no parecía ser significativa. Esto puede justificarse debido al tamaño de cada grupo ya que se están comparando 3207 de la institución 1710 del clúster 1, con 20 registros de la sede 1730 del clúster

En análisis para cada uno de los módulos por grupo de referencia otorgan un panorama de las fortalezas de los programas académicos. Destacándose Ingeniería en áreas de razonamiento cuantitativo, justificación que se haya en que estas ca-

rteras tienen una relación alta con modelos matemáticos, así como resolución de problemas de índole social. Otras de las carreras que sobresalen es la carrera de Medicina con puntajes superiores en áreas relacionadas con pensamiento crítico y comunicación escrita.

Las referencias cruzadas de todas las variables no se tuvieron en cuenta en esta investigación debida que en muchos casos se identificaban casos muy específicos, lo que no permitían hacer comparaciones estadísticas en relación a la base de datos. En este sentido, al trabajar con grupos más grandes, y por tanto, hacer comparaciones facilitó el ejercicio inferencial el cual servirá para realizar planes para mejoramiento continuo y la autoevaluación institucional.

Algunas limitaciones se encontraron en la capacidad de procesamiento de los computadores, aunque se puede entender que el número de relaciones a analizar eran de combinaciones muy grandes lo que excedía la capacidad de las máquinas.

Recomendaciones para futuras investigaciones son las de seguir ampliar la base de datos para cada una de las instituciones. Con esto se espera conocer si las significancias, como es el caso de la sede 1730 del clúster 3, cambian o se mantienen.

Otra de las recomendaciones para futuras investigaciones es la de realizar un análisis por bloques al azar para conocer el porcentaje de explicación en la variación general para cada uno de los puntajes en los módulos de las pruebas Saber Pro

Recibido: Junio de 2019

Aceptado: Febrero de 2020

Referencias

- Alfiani, A. P., Wulandari, F. A. et al. (2015), 'Mapping student's performance based on data mining approach (a case study)', *Agriculture and Agricultural Science Procedia* **3**, 173–177.
- Brown, S., White, S. & Power, N. (2017), 'Introductory anatomy and physiology in an undergraduate nursing curriculum', *Advances in physiology education* **41**(1), 56–61.
- Ferguson, T. D., Deephouse, D. L. & Ferguson, W. L. (2000), 'Do strategic groups differ in reputation?', *Strategic Management Journal* **21**(12), 1195–1214.
- Field, S., Kuczera, M. & Pont, B. (2007), 'No more failures: ten steps to equity in education. summary and policy recommendations'.
- for Economic Co-operation, O. & (OECD), D. (2010), *PISA 2009 results: what makes a school successful?: resources, policies and practices (vol. iv)*, OECD, Paris, France.

IBM Knowledge Center (2012), 'Log-likelihood distance (twostep clustering algorithms)', *Journal of Marketing Research* .
*www.ibm.com

OECD, Y. (2012), 'Equity and quality in education: Supporting disadvantaged students and schools'.

Punj, G. & Stewart, D. W. (1983), 'Cluster analysis in marketing research: Review and suggestions for application', *Journal of Marketing Research* **20**(2), 134–148.

Reardon, S. F. (2011), 'The widening academic achievement gap between the rich and the poor: New evidence and possible explanations', *Whither opportunity* pp. 91–116.

Sideridis, G. D., Mouzaki, A., Simos, P. & Protopapas, A. (2006), 'Classification of students with reading comprehension difficulties: The roles of motivation, affect, and psychopathology', *Learning Disability Quarterly* **29**(3), 159–180.

Zhao, J., Jin, L. & Shi, L. (2015), 'Mixture model selection via hierarchical bic', *Computational Statistics & Data Analysis* **88**, 139–153.