ARTÍCULO DE INVESTIGACIÓN / RESEARCH ARTICLE
http://dx.doi.org/10.14482/inde.37.2.1378

Caracterización de los estudiantes de una institución de educación superior mediante *big data**

Characterization of the students of a higher education institution through big data

Jorge Gabriel Hoyos Pineda**
Fredy Andrés Aponte-Novoa***

*Artículo de investigación científica y tecnológica, resultado del proyecto de investigación "Caracterización de los estudiantes de la USTA Tunja mediante *Big data*", con código SIS.2017-01, financiado por la universidad Santo Tomás Seccional Tunja.

**Universidad Pedagógica y Tecnológica de Colombia. Docente Escuela de Ingeniería de Sistemas y Computación Colombia. Magíster en Ciencias de la Información y las Comunicaciones. jorge.hoyos@uptc.edu.co

***Universidad Santo Tomás Seccional Tunja. Docente Facultad de Ingeniería de Sistemas. Magíster en Software Libre. fredy.aponte@usantoto.edu.co

Correspondencia: Jorge Gabriel Hoyos Pineda. Celular: 3004966267. Carrera 13 n.° 17-61 Apto. 202 Tunja (Boyacá)

Resumen

Este artículo presenta resultados de un proyecto de investigación cuyo objetivo principal consistió en caracterizar a los estudiantes de una institución de educación superior mediante el uso de técnicas y herramientas de *big data*. Esto con el propósito de dotar a la institución de una herramienta para apoyar la toma de decisiones relacionadas con la comunidad estudiantil. El proyecto se desarrolló en cinco fases: definición de la estrategia, captura y medición de los datos, análisis de los datos, generación de un informe de resultados y transformación del negocio. Como técnicas de análisis de datos se utilizaron el coeficiente de correlación de Pearson y el agrupamiento mediante k-means. Como resultado se obtuvo un inventario y caracterización de las fuentes de datos, y un modelo de análisis y procesamiento de la información que al ser aplicado genera una caracterización de una comunidad estudiantil.

Palabras clave: agrupamiento, análisis de datos, big data, caracterización de estudiantes, correlación de Pearson.

Abstract

This article presents results of a research project whose main objective was to characterize students of a higher education institution through the use of big data techniques and tools. This with the purpose of providing the institution with a tool to support decision-making related to the student community. The project was developed in five phases: strategy definition; data capture and measurement, data analysis, results report, and, business transformation. As data analysis techniques, the Pearson correlation coefficient and the k-means grouping were used. As a result, an inventory and characterization of the data sources was obtained, as well as a model of analysis and information processing, when it is applied, generates a characterization of a student community.

Keywords: big data, clustering, data analysis, Pearson correlation, students characterization.

Fecha de recepción: 11 de octubre de 2018
Fecha de aceptación: 17 de mayo de 2019

1. INTRODUCCIÓN

De acuerdo con los lineamientos sugeridos por el Consejo Nacional de Acreditación (CNA), uno de los factores más relevantes de la evaluación de la calidad de los programas de educación superior en Colombia es el relacionado con los estudiantes. Estos lineamientos establecen las características y aspectos que deben ser evaluados; entre los cuales se encuentran las "estrategias que garanticen la integración de los estudiantes a la institución en consideración a su heterogeneidad social y cultural", y la "deserción de estudiantes, análisis de causas y estrategias de permanencia en condiciones de calidad" [1].

Actualmente, la institución universitaria no cuenta con una herramienta que le permita hacer una caracterización de sus estudiantes, ya que la única información disponible para este propósito está compuesta por las estadísticas generadas a partir de la información registrada en su sistema de información académica. El contar con información más completa de los estudiantes, le permitiría a la institución mejorar el diseño de estrategias y acciones concretas por parte de diferentes dependencias académicas y administrativas que propendan por responder de forma oportuna y pertinente a las necesidades de la comunidad estudiantil.

Lo anterior llevó a considerar la oportunidad de utilizar técnicas y herramientas de *big data* para recolectar y analizar la información acerca de los estudiantes, información que actualmente reposa en bases de datos y archivos físicos y digitales.

El *big data* puede ser definido como la inteligencia colectiva generada y compartida en un entorno tecnológico, en el que prácticamente cualquier cosa puede ser transformada en datos a través de procesos de documentación, medición y captura digital [2].

Comúnmente el *big data* se define con las tres V: Volumen, Velocidad y Variedad. Estos tres términos están referidos a los activos de información que se caracterizan por su alto volumen, alta velocidad de generación y alta variedad, dada la diversidad de formatos utilizados. Estos activos de información requieren formas de procesamiento innovadoras y eficientes que permitan mejorar el conocimiento, la toma de decisiones y la automatización de procesos [3]. Otros autores van más allá y hablan hasta de 6 V: Volumen, Velocidad, Veracidad, Variedad, Verificación y Valor, y plantean que además de la información, el concepto de *big data* abarca las técnicas y tecnologías que hacen posible la captura, almacenamiento, distribución, administración y análisis de grandes conjuntos de datos con estructuras diversas [4].

En el campo de la educación, el uso del análisis de datos puede tener diferentes objetivos, dependiendo del grupo de interés de los usuarios y los participantes involucrados en el proceso (estudiantes, docentes, instituciones). En el caso de las instituciones, el objetivo es mejorar y hacer más eficientes los procesos de toma de

decisiones [5]. El desarrollo que han tenido las técnicas y tecnologías del *big data* ha generado nuevas oportunidades para explorar el proceso de aprendizaje de los estudiantes y formas eficientes de mejorar dicho proceso [6].

Con ayuda del *big data* los administradores de las instituciones educativas pueden mejorar la organización de los recursos de aprendizaje y formular la dirección de las reformas educativas y hacer mediciones, mientras que los directivos encargados de formular las políticas pueden utilizarlo para definir políticas con un mayor respaldo de evidencia. Entre mejor se entienda a los estudiantes, mejores serán los logros que ellos mismos puedan alcanzar [5], [7], [8]. En [7] se resaltan las grandes potencialidades que tiene el uso de *big data* en diferentes aspectos de la educación superior, como son las tendencias de admisión, los problemas de retención y promoción de estudiantes, análisis de la inversión y el impacto de la investigación, optimización de la infraestructura, entre otros.

En el desarrollo del proyecto se utilizó la metodología SMART, compuesta por cinco fases: definición de la estrategia; captura y medición de los datos; análisis de los datos; generación de un informe de resultados y transformación del negocio.

2. METODOLOGÍA

Basados en la propuesta de [8], en el proceso de *big data* se consideraron las cinco etapas contempladas en la metodología SMART: definición de la estrategia; captura y medición de los datos; análisis de los datos; generación de un informe de resultados y transformación del negocio. Estas fases guardan correspondencia con las definidas en las metodologías clásicas de la minería de datos. A manera de ilustración, en la tabla 1 se muestra la equivalencia de cada una de las fases de la metodología SMART con la metodología CRISP-DM:

Tabla 1. Equivalencia entre fases de las metodologías SMART y CRISP-DM

SMART	CRISP-DM
Definición de la estrategia	Entendimiento del negocio
Captura y medición de los datos	Entendimiento de los datos
	Preparación de los datos
Análisis de los datos	Modelado
	Evaluación
Generación de un informe de resultados	Despliegue
Transformación del negocio	

Fuente: Los autores.

A continuación se describe cada una de las fases desarrolladas.

Etapa 1. Definición de la estrategia

El equipo investigador definió que debería centrarse en la información del estudiante, la social (ciudad e institución de origen, domicilio, conformación grupo familiar, intereses, aficiones, pertenencia a grupos), la académica (resultado proceso de admisión, promedios, permanencia, repitencia), con el fin de obtener un mejor conocimiento de la población estudiantil, que le permitiera a la institución reorientar políticas y estrategias dirigidas a esta parte de la comunidad universitaria.

De esta forma, y de acuerdo con la revisión de las fuentes, se escogieron 21 variables correspondientes a la información tanto académica como socioeconómica de cada estudiante matriculado.

Etapa 2. Captura y medición de los datos

En primer lugar se analizó la estructura y contenido de las fuentes de datos para identificar el tipo de información almacenada, la forma de acceso y su disponibilidad, para posteriormente extraer la información de las fuentes y prepararla para la fase de análisis.

Definición de las variables de interés. El equipo investigador procedió a definir las variables de interés para el estudio propuesto y a organizarlas en tres categorías: demográficas, desempeño e institucionales. En la tabla 2 se puede ver el detalle de esta definición.

Inventario de fuentes. Una vez identificadas las variables de interés, se procedió a establecer las posibles fuentes de datos que podrían ser utilizadas: Sistema Académico, Oficina de Admisiones, historias académicas. En la tabla 3 se describen las posibles fuentes de datos que podrían ser utilizadas en la investigación.

Caracterización de las fuentes. Una vez identificada las fuentes de datos, se procede a caracterizarlas de acuerdo con cinco criterios: forma de almacenamiento, niveles de acceso, estructura, completitud y veracidad, y vacíos de información.

Fuentes por utilizar. Identificadas y caracterizadas las fuentes de datos, se procede a definir la fuente de información a utilizar para cada una de las variables propuestas. De las 21 variables seleccionadas, se establece el Sistema Académico para 11 de ellas, y la Oficina de admisiones para las 10 restantes, como se relaciona a continuación. Oficina de Admisiones: Estrato, Tipo de vivienda, Género, Domicilio, Ciudad de origen, IE de origen, Núcleo Familiar, Resultados saber 11 y Puntaje entrevista. Sistema Académico: Fecha de nacimiento, Programa académico, Periodo de Inicio,

Nivel o semestre actual, Promedio acumulado, Número de asignaturas aprobadas, Número de asignaturas perdidas, Número de matrículas realizadas, Código, Identificación, Nombre y Correo personal.

Tabla 2. Variables de interés

Demográficas	Desempeño (registros académicos)	Institucionales (contexto y de apoyo)
Estrato Tipo de vivienda Género Domicilio Ciudad de origen IE de origen Núcleo familiar Ingreso familiar Fecha de nacimiento	Programa académico Período de inicio Nivel o semestre actual Promedio acumulado Número de asignaturas aprobadas Número de asignaturas pérdidas Número de matrículas realizadas	Código Identificación Nombre Correo electrónico personal Resultados saber 11 Puntaje entrevista

Fuente: Los autores.

Tabla 3. Inventario de fuentes de información

Fuente	Descripción
Sistema Académico	Sistema de información que mantiene la información académica de los estudiantes generada en toda su vida universitaria.
Oficina de Admisiones	Información socioeconómica de los aspirantes y nuevos estudiantes vinculados a la institución.
Historias académicas	Archivo físico que reposa en cada unidad académica que contiene toda la documentación relacionada con el proceso del estudiante.

Fuente: Los autores.

Preparación de los datos. Se obtuvieron dos archivos con la información relativa a los estudiantes matriculados para el primer semestre de 2017. El primer archivo cuenta con 3908 registros con los siguientes 23 campos: código alumno, numero

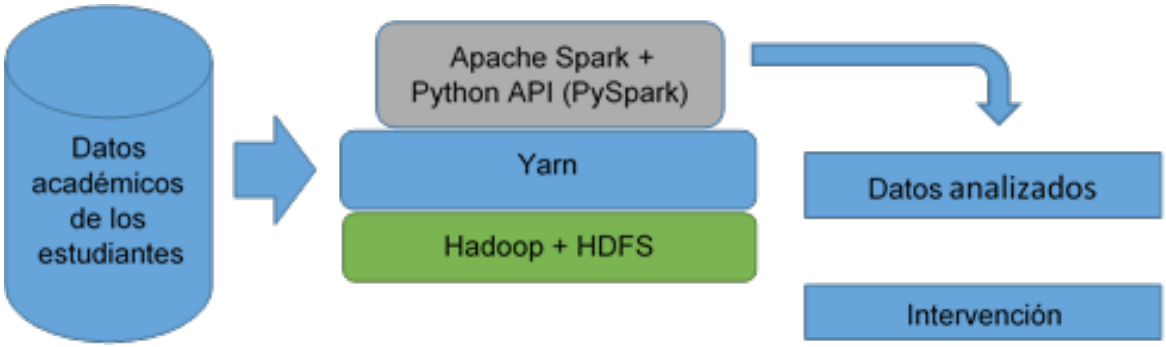
identificación, primer apellido, segundo apellido, primer nombre, segundo nombre, e-mail personal, nombre unidad, fecha nacimiento, periodo ingreso, número nivel cursado, género estudiante, dirección residencia, municipio de origen, jornada, dirección e-mail institucional, edad en años, promedio acumulado, número estrato socioeconómico, colegio, cantidad asignaturas aprobadas, cantidad asignaturas perdidas, número de matrículas.

El segundo archivo se compone de 12 957 registros, con los datos de los aspirantes a ingreso desde el segundo semestre de 2014 al segundo de 2017, y tiene los siguientes campos: código periodo, nombre unidad, numero identificación, nombre largo, código programa, resultados saber 11 anterior (Biología, Matemáticas, Filosofía, Física, Historia, Química, Lenguaje, Geografía, Inglés, Ciencias Sociales y puntaje total), y resultados saber 11 nuevo (Lectura Crítica, Sociales y Ciudadanas, Ciencias Naturales, Razonamiento Cuantitativo y Competencias Ciudadanas).

Posteriormente se realiza un proceso de limpieza de datos. En el primer archivo se unen los campos primer apellido, segundo apellido, primer nombre, segundo nombre para generar el campo nombre completo, el cual se usó para cruzar con el archivo 2. También se elimina la columna "DIR_RESIDENCIA", pues no aporta información de interés. Del segundo archivo se eliminan las columnas NOM_UNIDAD y COD_PROG_OPC_UNO; la primera de estas por estar duplicada con el primer archivo; la segunda, por no aportar nada al análisis. Adicionalmente, se eliminaron espacios innecesarios, tildes y caracteres especiales. Finalmente se cruzan los dos archivos mediante el campo nombre completo, y se obtiene un archivo con los datos limpios.

Etapa 3. Análisis de los datos

Esta etapa resultó la más compleja, ya que abarcó dos grandes temas: la infraestructura tecnológica para soportar el proceso de *big data* y la selección de las técnicas de análisis de datos que serían utilizadas. En el primer aspecto, la infraestructura tecnológica, a partir de la revisión bibliográfica se optó por utilizar un conjunto de herramientas de uso libre que ya hubieran sido utilizadas y probadas en proyectos similares y que correspondieran a una configuración factible de reproducir de acuerdo con las restricciones de tipo técnico del proyecto. Es así como se diseñó y configuró un clúster con un equipo maestro y dos equipos esclavos, que sirvieran de base para la instalación de Hadoop con el sistema de archivos HDFS y el gestor de recursos YARN, base sobre la cual funciona Spark, que es el framework que proporciona las librerías especializadas en el análisis de datos. Respecto a las técnicas de análisis, se seleccionaron el coeficiente de correlación de Pearson y el agrupamiento mediante k-means. En la figura 1 se muestra un esquema del modelo utilizado.



Fuente: Los autores. Adaptado de [9].

Figura 1. Modelo de análisis y procesamiento utilizado

Etapa 4. Generación de un informe de resultados

En esta etapa se generó un informe de los resultados obtenidos a partir de la aplicación de las técnicas de análisis seleccionadas. En primer lugar, se generaron gráficos estadísticos que permiten tener una primera aproximación a la información contenida en los datos, y posteriormente se realizaron los análisis correspondientes. En la sección de resultados se describen en detalle los valores obtenidos.

Etapa 5. Transformación del negocio

Una vez socializados los resultados, serán las unidades académicas y administrativas las que podrán reorientar algunas de sus estrategias y actividades. Sin embargo, el equipo que desarrolla el proyecto estará en capacidad de identificar nuevas oportunidades de aplicación del modelo en otras áreas de la institución.

3. RESULTADOS Y DISCUSIÓN

En cuanto a la distribución de los estudiantes en los diferentes semestres, se observa que en los semestres impares la población de estudiantes está cercana a duplicar el número de estudiantes de los semestres pares, y que esta tendencia se mantiene a lo largo de los diez semestres académicos que conforman la mayoría de los programas académicos, y que con la progresión del número de semestre se va disminuyendo la población de cada cohorte. Lo anterior se explica por el hecho de que en razón del calendario académico A (de febrero a noviembre) que manejan las instituciones educativas en la región, las matrículas para el primer semestre del año son mucho mayores que las correspondientes al segundo semestre, comportamiento que resulta normal en esta parte del país. En la tabla 4 se puede ver la valoración de cada una de características analizadas.

Tabla 4. Caracterización de la población estudiantil

Característica	Valor
Hombres	54%
Mujeres	46%
Proviene de institución oficial	54%
Proviene de institución privada	46%
Jornada diurna	94%
Jornada nocturna	6%
Entre 15 y 20 años	45%
Entre 21 y 25 años	46%
Entre 26 y 30 años	7%
Entre 31 y 40 años	2%
Estrato socioeconómico 1	7%
Estrato socioeconómico 2	44%
Estrato socioeconómico 3	37%
Estrato socioeconómico 4	9%
Estrato socioeconómico 5	3%
Ciudad de origen igual a sede institución	41%
Ciudad de origen otros municipios del departamento	42%
Ciudad de origen otras zonas del país	17%
Promedio académico entre 4,1 y 5,0	19%
Promedio académico entre 3,1 y 4,0	75%
Promedio académico entre 2,1 y 3,0	4%
Promedio académico entre 0,0 y 2,0	2%

Fuente: Los autores.

Respecto a los datos presentados en la tabla 4 se resalta el hecho de que un 88 % de la población estudiantil se concentra en los estratos 1, 2 y 3.

Al analizar el número de matrículas registradas para un mismo estudiante, se encontró que un número pequeño presenta más de once matrículas, que es considerado el número promedio de semestres para terminación de materias, y que en el nivel más extremo aparecen cinco estudiantes con 16 matrículas.

Una vez realizada la caracterización individual de las variables de mayor interés, se procedió a aplicar técnicas de análisis para buscar la relación existente entre las diferentes variables involucradas.

Un primer ejercicio consistió en generar la matriz de coeficientes de correlación de Pearson, con el fin de identificar la existencia de variables correlacionadas y el grado de correlación entre las mismas. Este coeficiente mide el grado o magnitud de la relación entre diferentes variables, partiendo del supuesto de que existe una relación lineal entre las mismas. El valor del coeficiente varía entre -1 y 1. La magnitud de la relación está dada por el valor absoluto, mientras que el signo refleja la dirección de la relación. En este sentido, un valor de 1 representa una relación perfecta positiva, que indica que la relación entre las variables es fuerte y directa, es decir que, al aumentar el valor de una variable, la otra también aumenta en forma proporcional. De la misma forma, un valor de -1 representa una relación perfecta negativa, lo cual indica que la relación es fuerte e indirecta. Un valor de 0 indicaría que no existe relación entre las variables involucradas, y los valores intermedios indicarían una relación fuerte o débil de acuerdo con su cercanía con los extremos, -1 o 1 [10]. Una vez generada la matriz de coeficientes, se encontró que las únicas correlaciones fuertes que se aprecian son entre el número de niveles cursados, el número de asignaturas aprobadas y el número de matrículas, relación que en principio parece obvia. Por otro lado, se evidencia que en el caso de la edad, el estrato socioeconómico y el número de asignaturas pérdidas no presentan una relación fuerte con las demás variables analizadas. Los valores obtenidos se muestran en la figura 2.

NUM_NIV_CURSA	AÑOS	NUM_EST_ECONOMICO	ASIG_APROB	ASIG_PERD	NUM_MATRICULAS	BIOLOGIA	MATEMATICAS	FILOSOFIA	FISICA	HISTORIA	QUIMICA	LENGUAJE	GEOGRAFIA	INGLES	ENTREVISTA	CIENCIAS_SOCIALES	ICFES_ANTERIOR	LECTURA_CRITICA	SOCIALES_Y_CIUDADANAS	CIENCIAS_NATURALES
1,00	0,43	-0,43	0,94	0,23	0,86	-0,17	-0,23	0,30	0,36	-0,15	0,23	0,65	-0,15	-0,69	0,34	0,29	0,12	-0,61	-0,61	-0,61
0,43	1,00	-0,17	0,43	0,33	0,44	0,19	-0,25	0,35	0,31	0,19	0,27	0,49	0,17	-0,46	0,08	0,28	0,22	-0,55	-0,55	-0,55
-0,43	-0,17	1,00	-0,43	-0,23	-0,44	0,15	0,10	-0,11	-0,25	0,15	-0,17	-0,38	0,14	0,43	-0,16	-0,15	-0,05	0,35	0,35	0,35
0,94	0,43	-0,43	1,00	0,23	0,87	-0,13	-0,25	0,38	0,33	-0,12	0,21	0,68	-0,13	-0,65	0,37	0,38	0,13	-0,60	-0,59	-0,60
0,23	0,33	-0,23	0,23	1,00	0,52	-0,06	-0,12	0,04	0,22	-0,05	0,19	0,27	-0,06	-0,31	0,10	0,06	-0,01	-0,28	-0,28	-0,28
0,86	0,44	-0,44	0,87	0,52	1,00	-0,17	-0,20	0,21	0,42	-0,15	0,29	0,62	-0,15	-0,63	0,34	0,20	0,01	-0,56	-0,56	-0,56
-0,17	0,19	0,15	-0,13	-0,06	-0,17	1,00	0,09	0,44	0,43	0,90	0,52	0,30	0,86	0,21	-0,08	0,45	-0,07	-0,27	-0,27	-0,27
-0,23	-0,25	0,10	-0,25	-0,12	-0,20	0,09	1,00	-0,34	0,24	0,07	0,19	-0,19	0,09	0,51	-0,05	-0,34	-0,03	0,47	0,47	0,47
0,30	0,35	-0,11	0,38	0,04	0,21	0,44	-0,34	1,00	0,13	0,43	0,04	0,65	0,39	-0,33	0,15	0,88	0,21	-0,56	-0,56	-0,56
0,36	0,31	-0,25	0,33	0,22	0,42	0,43	0,24	0,13	1,00	0,40	0,79	0,65	0,36	-0,36	0,16	0,19	-0,14	-0,58	-0,58	-0,58
-0,15	0,19	0,15	-0,12	-0,05	-0,15	0,90	0,07	0,43	0,40	1,00	0,49	0,28	0,94	0,19	-0,06	0,41	-0,06	-0,25	-0,25	-0,25
0,23	0,27	-0,17	0,21	0,19	0,29	0,52	0,19	0,04	0,79	0,49	1,00	0,53	0,46	-0,20	0,11	0,04	-0,11	-0,47	-0,47	-0,47
0,65	0,49	-0,38	0,68	0,27	0,62	0,30	-0,19	0,65	0,65	0,28	0,53	1,00	0,25	-0,69	0,29	0,62	0,13	-0,85	-0,85	-0,85
-0,15	0,17	0,14	-0,13	-0,06	-0,15	0,86	0,09	0,39	0,36	0,94	0,46	0,25	1,00	0,17	-0,09	0,39	-0,06	-0,21	-0,21	-0,21
-0,69	-0,46	0,43	-0,65	-0,31	-0,63	0,21	0,51	-0,33	-0,36	0,19	-0,20	-0,69	0,17	1,00	-0,23	-0,29	-0,18	0,84	0,84	0,84
0,34	0,08	-0,16	0,37	0,10	0,34	-0,08	-0,05	0,15	0,16	-0,06	0,11	0,29	-0,09	-0,23	1,00	0,16	0,04	-0,21	-0,21	-0,21
0,29	0,28	-0,15	0,38	0,06	0,20	0,45	-0,34	0,88	0,19	0,41	0,04	0,62	0,39	-0,29	0,16	1,00	0,17	-0,53	-0,53	-0,53
0,12	0,22	-0,05	0,13	-0,01	0,01	-0,07	-0,03	0,21	-0,14	-0,06	-0,11	0,13	-0,06	-0,18	0,04	0,17	1,00	-0,15	-0,15	-0,15
-0,61	-0,55	0,35	-0,60	-0,28	-0,56	-0,27	0,47	-0,56	-0,58	-0,25	-0,47	-0,85	-0,21	0,84	-0,21	-0,53	-0,15	1,00	0,99	0,99
-0,61	-0,55	0,35	-0,59	-0,28	-0,56	-0,27	0,47	-0,56	-0,58	-0,25	-0,47	-0,85	-0,21	0,84	-0,21	-0,53	-0,15	0,99	1,00	0,99
-0,61	-0,55	0,35	-0,60	-0,28	-0,56	-0,27	0,47	-0,56	-0,58	-0,25	-0,47	-0,85	-0,21	0,84	-0,21	-0,53	-0,15	0,99	0,99	1,00

Fuente: Los autores.

Figura 2. Matriz de coeficientes de correlación de Pearson

Con el fin de identificar conjuntos de estudiantes similares, basado en características demográficas y la información académica, se optó por utilizar el agrupamiento en clústeres mediante k-means, ya que resulta una técnica útil para administrar y organizar grandes volúmenes de datos [11]. Está técnica de agrupamiento mediante un proceso iterativo ajusta los centroides o valores centrales que determinan la pertenencia o no a un grupo en particular de acuerdo con la semejanza de sus características [12]. Fundamentalmente busca minimizar las diferencias entre individuos de un mismo grupo y maximizar esas diferencias respecto a individuos de grupos diferentes. Los resultados de este proceso se muestran en la figura 3.

CLASE	0	1	2	3	4
INDIVIDUOS	833	229	1294	106	1443
NUM_NIV_CURSA	3,04	3,70	3,52	6,72	7,01
AÑOS	19,29	22,97	21,04	24,26	22,55
NUM_EST_ECONOMICO	2,56	2,52	2,22	1,69	1,23
ASIG_APROB	22,59	28,09	24,52	48,66	49,92
ASIG_PERD	1,42	1,86	2,05	2,44	4,61
NUM_MATRICULAS	3,04	3,60	3,43	5,30	7,58
BIOLOGIA	0,00	48,01	6,83	0,00	0,24
MATEMATICAS	58,21	51,35	47,23	45,20	37,28
FILOSOFIA	0,00	44,84	11,70	41,91	22,22
FISICA	0,00	47,90	12,14	0,44	29,45
HISTORIA	0,00	45,08	5,73	0,00	0,13
QUIMICA	0,00	47,92	9,62	0,40	19,95
LENGUAJE	0,00	50,04	18,04	46,71	48,28
GEOGRAFIA	0,00	44,64	6,59	0,00	0,00
INGLES	56,64	44,38	35,51	0,00	0,02
ENTREVISTA	315,98	318,56	0,04	337,74	323,29
CIENCIAS_SOCIALES	0,00	46,09	10,74	37,48	22,84
ICFES_ANTERIOR	0,00	0,00	7,45	251,07	0,07
LECTURA_CRITICA	56,75	0,00	30,39	0,00	0,00
SOCIALES_Y_CIUDADANA	57,68	0,00	31,06	0,00	0,00
CIENCIAS_NATURALES	57,49	0,00	31,06	0,00	0,00

Fuente: Los autores.

Figura 3. Resultados de la aplicación del algoritmo k-means para cinco grupos

Uno de los parámetros de entrada del algoritmo k-means es el número de grupos en los que se quiere agrupar la población total. Resulta común que el algoritmo se ejecute tres o cuatro veces para tres, cuatro o cinco grupos [11]. En este caso, el algoritmo fue ejecutado para 5 grupos y se obtuvo la siguiente distribución: grupo 0, conformado por 833 individuos (21 %), correspondiente a estudiantes en promedio de 19 años y tres semestres cursados; grupo 1 conformado por 229 individuos (6 %), correspondiente a estudiantes en promedio de 23 años y cuatro semestres cursados; grupo 2, conformado por 1294 individuos (33 %), correspondiente a estudiantes en promedio de 21 años y tres semestres cursados; grupo 3, conformado por 106 individuos (3 %), correspondiente a estudiantes en promedio de 24 años y cinco semestres cursados; y grupo 4, conformado por 1443 individuos (37 %), correspondiente a estudiantes en promedio de 23 años y ocho semestres cursados. Este último grupo podría considerarse como el de mejor desempeño, dada la consistencia entre la edad y en nivel académico. Al realizar este agrupamiento se obtiene los centroides de cada uno de los grupos, los cuales proporcionan información acerca de los valores medios para cada una de las variables correspondientes a cada individuo que fue asignado a determinado grupo. Se realizaron pruebas adicionales que generaron 4 grupos, y se obtuvieron resultados similares.

En un ejercicio posterior se generó la matriz de coeficientes de Pearson para un grupo en particular de los generados por k-means. En este caso se reafirma que aunque solo se está examinando el primero de los agrupamientos, las únicas correlaciones fuertes que se aprecian son entre en número de niveles cursados, el número de asignaturas aprobadas y el número de matrículas, relación que en principio parece obvia.

4. CONCLUSIONES

En el desarrollo del proyecto se presentaron dos situaciones relacionadas con el acceso a la información que pudieron limitar en gran medida los alcances del proyecto y el impacto del mismo. En primer lugar, la calidad y disponibilidad de la información académica y socioeconómica de los estudiantes que reposa en los sistemas de información de la institución universitaria. La única fuente completa y confiable con que se pudo contar fue el Sistema Académico, y se destaca que la cantidad de información es limitada, mantiene problemas de completitud y no se puede acceder de forma directa, sino a través de archivos planos. El segundo tema tiene que ver con el acceso a la información de la red social Facebook, información que es restringida por políticas de esa empresa, y a la cual se puede acceder en forma parcial, pero comprando la información, aspecto que no había sido contemplado en el presupuesto del proyecto.

Una vez generada la matriz de coeficientes de Pearson, se encontró que las únicas correlaciones fuertes entre variables están dadas entre el número de niveles cursados, el número de asignaturas aprobadas y el número de matrículas, relación que por su naturaleza parece obvia. Por otro lado, se evidencia que en el caso de la edad, el estrato socioeconómico, y el número de asignaturas pérdidas no presentan una relación fuerte con las demás variables analizadas, lo cual significa que el comportamiento de estas variables no afecta el comportamiento de las demás.

Soluciones como la que se plantea se convierten en herramientas de gran utilidad para una institución de educación superior, teniendo en cuenta que atiende necesidades de información que resulta muy valiosa para apoyar la toma de decisiones en aspectos tan importantes para este tipo de organizaciones como es el desarrollo integral de su comunidad estudiantil.

REFERENCIAS

[1] Consejo Nacional de Acreditación, Lineamientos para la acreditación institucional, 2015.

[2] U. Sivarajah, M. Kamal, Z. Irani y V. Weerakkody, «Critical analysis of Big Data challenges anda analytical methods», *Journal of Business Research*, Vol.70, p. 263-286, 2017. doi: 10.1016/j.jbusres.2016.08.001

[3] Gartner, Inc, "Gartner It Glossary", 2016. Available: http://www.gartner.com/it-glossary/big-data/.

[4] B. Daniel, «Big Data and analytics in higher education: Oportunities and challenges», *British Journal of Educational Technology*, Vol.46 Num.5, pp. 904-920, 2015. doi: 10.1111/bjet.12230

[5] X. Yu & S. Wu, "Typical Applications of Big Data in Education", en *International Conference of Educational Innovation through Technology*, 2015. doi: 10.1109/EITT.2015.29

[6] L. Calvet y Á. Juan, «Educational Data Mining and Learning Analytics: differences, similarities, and time evolution», Universities and Knowledge Society Journal, Vol.12 Num.3, pp. 98-112, 2015. doi: 10.7238/rusc.v12i3.2515

[7] V. Koon Ong, "Big Data and its Research Implications for Higher Education: Cases from UK Higher", en *IIAI 4th International Congress on Advanced Applied Informatics*, London, 2015. doi: 10.1109/IIAI-AAI.2015.178

[8] E. Luría et al., "Crossing the Cash to Big Data: Early Detection of At-Risk Students in a Cluster Computing Environment", en *7a Conferencia Internacional en analítica del aprendizaje y el conocimiento*, Vancouver, 2017.

[9} B. Marr, «Big data in practice», *Journal of Business Research*, pp. 366-378. citado en Sharif, A. M., Shah, N., & Irani, Z. (2017). "*Big data in an HR context: Exploring organizational change readiness, employee attitudes and behaviors*", *Journal of Business Research*, Vol.70, 2015. doi: 10.1016/j.jbusres.2016.08.010

[10] A. Alver & L. Altas, "Characterization and electrocoagulative treatment of landfill leachates: A statistical approach", *Process Safety and Environmental Protection,* pp. 102-111, 2017. doi: 10.1016/j.psep.2017.04.021

[11] P. Antonenko, S. Toy & D. Niederhauser, "Using cluster analysis for data mining in educational technology research", *Education Tech Research Dev*, vol. 60, pp. 383-398. doi 10.1007/s11423-012-9235-8, 2012.

[12] A. Kathuria et al., "Classifiying the user intent of web queries using k-means clustering", *Internet Research*, pp. 563-581, 2010. doi: 10.1108/10662241011084112