Implementación de un sistema de análisis de datos en la deserción estudiantil utilizando técnicas de Big Data para facilitar la estructuración de planes de mejoramiento de la Universidad Mariana

José Javier Villalba Romero
Robinson Andrés Jiménez Toledo
Docentes del Programa de Ingeniería de Sistemas
Universidad Mariana

Luis Eduardo Paredes
Estudiante del Programa de Ingeniería de Sistemas
Universidad Mariana

Fuente: pixabay.

El Ministerio de Educación Superior Colombiano (Ministerio de Educación, 2009) afirma que "la deserción estudiantil en las instituciones de educación superior se concibe como un fenómeno que afecta significativamente la población estudiantil en el país, es una problemática que no distingue entre estrato social, ideologías y religión". En el departamento de Nariño existe un agravante en las instituciones de educación superior, según la Secretaria de Educación (Secreteria de Educacion de Nariño, 2014) explica que "es la deserción estudiantil uno de los principales factores que incide en el desarrollo de la educación". Es por eso que, en la mayoría de las universidades en Colombia, especialmente en las universidades del departamento de Nariño, este fenómeno se ha hecho muy complejo a la hora de afrontarlo.

Es fundamental identificar cuáles son los factores que afectan directamente en la educación superior, es por eso que, la importancia de entender el significado de la deserción estudiantil universitaria, y cuáles son las consecuencias que pueden perjudicar tanto al estudiante como a la universidad.

La deserción estudiantil universitaria es una de las principales problemáticas en el país, en donde existe un gran infortunio en los jóvenes que deciden ingresar al sistema de educación superior, según las estadísticas arrojadas por el Sistema para la Prevención de la Deserción de la Educación Superior (SPADIES) (Ministerio de Educación Superior Colombiano, 2008), el 48,47% de los estudiantes que ingresaron a la educación superior en el primer semestre de 2000, no alcanzaron décimo semestre; mientras que el 57.2% de los que ingresaron en el primer semestre de 2008 no lo hicieron.

Se debe asimilar o diferenciar lo que significa deserción estudiantil universitaria, según Correa (Universidad EAFIT, 1999), la deserción estudiantil entendida no sólo como el abandono definitivo de las aulas de clase, sino como el abandono de la formación académica, independientemente de las condiciones y modalidades de preespecialidad, es decisión personal del sujeto y no obedece a un retiro académico forzoso (por el no éxito del estudiante en el rendimiento académico, como es el caso de expulsión por bajo promedio académico), o el retiro por asuntos disciplinares. Diríase entonces, que la deserción es opción del estudiante, influenciando positiva o negativamente por circunstancias internas o externas.

El problema de la deserción estudiantil en la Universidad Mariana es un tema que se debe abordar lo más pronto posible, porque se ha evidenciado en otras instituciones que la deserción estudiantil universitaria es uno de los problemas más propensos a deteriorar el desarrollo de la comunidad estudiantil en general, se ha detectado que los estudiantes que estaban en sus inicios como universitarios, deciden cambiar de rumbo o simplemente desertan en ocasiones por motivos desconocidos, ya que no se tiene un registro específico acerca de este grave flagelo; este problema se podría atacar o contrarrestar si los estudiantes tuvieran una posibilidad de tomar una serie de opciones que impidan que deserten.

La carencia de sistema de análisis y predicción de datos que provea de información precisa y oportuna a las facultades en el pronóstico del comportamiento de la deserción estudiantil, hace que la tomar decisiones se haga de manera empírica y no soportada en los procesos mediados por TI. Para desarrollo de la presente investigación, se implementó un sistema de análisis y predicción de datos en la deserción estudiantil de la Universidad Mariana, basada en técnicas de Big Data, que facilitaría la estructuración de planes de mejoramiento.

De igual manera, se identificó las características en cuanto a volumen, variedad, veracidad y velocidad de los datos que se manejan de los estudiantes desertados de la Universidad Mariana Pasto. Además, se reconoció las características y técnicas de Hadoop en el uso de Big Data para sistemas de predicción y pronósticos; por último, se implementó un clúster computacional basado en Hadoop que soporta el sistema de pronóstico de deserción estudiantil. La presente investigación se desarrolló en la Universidad Mariana, en su sede principal sede de la ciudad de Pasto (Universidad Mariana, 1970). La Universidad Mariana tiene varias facultades con sus respectivos programas, los cuales son objeto de estudio para esta investigación.

La presente investigación tomó como base los procesos de investigación que se describen a continuación:

Paradigma: la investigación se ubica dentro de un paradigma cuantitativo deductivo, ya que según Sampieri (Hernández, Fernández y Baptista, 2003), el paradigma cuantitativo utiliza la recolección y el análisis de datos para contestar preguntas de investigación y probar hipótesis establecidas previamente, y confía en la medición numérica, el conteo y frecuentemente, en el uso de la estadística para establecer con exactitud patrones de comportamiento en una población, y para efectos de la investigación se empleará el paradigma cuantitativo, ya que se hará procesos numéricos y estadísticos para analizar y evaluar resultados que lleven a una posible aceptación de la hipótesis planteada.

Enfoque: el enfoque de la investigación es empírico analítico, ya que según Habermas (Jürgen, 1977), busca la explicación, la determinación de causas y efectos cuantitativamente comprobables y repetibles en contextos diversos con variables de control. La realidad se desagrega por variables cuantificables, se buscan regularidades que permitan proposiciones. Su interés es técnico, ambiciona predecir y controlar los hechos que estudia para modificarlos. Este enfoque aplica para esta investigación, ya que se plantea predecir y pronosticar hechos que se estudian para determinar posibles soluciones, para la investigación se usarán variables cuantificables, lo cual supone buscar la explicación, determinación de causas y efectos cuantitativamente comprobables.

Tipo de investigación: la investigación es aplicada, según la Universidad Nacional (Universidad Nacional a Distancia, 2014). La investigación científica aplicada se propone transformar el conocimiento 'puro' en conocimiento útil. Tiene por finalidad la búsqueda y consolidación del saber y la aplicación de los conocimientos para el enriquecimiento del acervo cultural y científico, así como la producción de tecnología al servicio del desarrollo integral de las naciones. La investigación aplicada puede ser fundamental o tecnológica.

En este caso, la investigación aplicada es tecnológica, se entiende como aquella que genera conocimientos o métodos dirigidos al sector productivo de bienes y servicios, ya sea con el fin de mejorarlo y hacerlo más eficiente, o con el fin de obtener productos nuevos y competitivos en dicho sector. En la investigación se implementó un clúster computacional y se desarrolló una herramienta software para el procesamiento de los datos de la deserción estudiantil.

La línea de investigación para esta investigación estuvo enfocada en el área de Ingeniería, Informática y Computación. La población objeto de investigación fueron todos los datos de los estudiantes de la Universidad Mariana de Pasto, y se tomaron como muestra únicamente los datos de los estudiantes desertores entre los años 2006 a 2016, quienes representaron a la población objeto de estudio.

El proceso de investigación se especificó de acuerdo a la siguiente Tabla 1.

Tabla 1. *Procesos de investigación*

Objetivos específicos	Fuente	Técnica de recolección	Instrumento	Técnica de Procesamiento	Resultado
Identificar las características de los datos que se manejan de los estudiantes desertados de la Universidad Mariana	Base de datos de los estudiantes desertores de la Universidad Mariana	Observación directa	Libreta de apuntes, lista de chequeo	Síntesis	Informe detallado de la variabilidad de los datos de estudiantes desertados de la Universidad Mariana

Reconocer las características y técnicas de Hadoop en el uso de Big Data para sistemas de predicción y pronósticos	Google académico, libros, revistas, artículos científicos.	Revisión documental	Fichas de revisión	Análisis documental	Informe detallado de las características y técnicas de Hadoop en el uso Big Data
Implementar un clúster computacional basado en Hadoop que soporta el sistema de pronóstico de deserción estudiantil	Google académico, tutoriales, artículos científicos	Revisión documental	Fichas de revisión	Análisis documental	Clúster computacional para análisis de datos implementado

Para la investigación se usó variables volumen de los datos, variabilidad de los datos y la eficiencia del sistema.

Tabla 2. *Variables e hipótesis*

Variable	Descripción	Tipo	Indicador	Naturaleza	Fuente
Volumen de los datos	Esta variable determina el tamaño de los datos de estudiantes desertados de la Universidad Mariana	I	Cantidad o peso de los datos de la base de datos	Cuantitativa	Base de datos de los estudiantes desertores la Universidad Mariana.
Variabilidad de los datos	Esta variable determina el tipo de datos, su formato y su estructura.	D	Nivel de variabilidad de los datos	Cuantitativa	
Eficiencia del sistema	La variable evalúa el nivel de rendimiento del sistema en cuanto a procesamiento.	D	Tiempos de respuesta del sistema de análisis de datos deserción estudiantil.	Cuantitativa	Clúster computacional y el sistema de análisis de datos

Independiente I, Dependiente D.

Resultados parciales técnicos de la investigación

Identificación de los datos de cada estudiante necesarios para trabajar con ellos el análisis de datos de los cuales se solicitaron a la oficina de informática de la universidad Mariana.

A continuación, se muestra una imagen (Ver figura 1, figura 2) con todos los datos necesarios que se solicitó a la oficina de Tecno Smart para que se entregara una base de datos completa en cuanto a los registros solicitados.

Datos Estudiante cada programa -cada facultada											
Codigo	Nombre	Programa	Fecha_inicio	Fecha_final	Modalidad	Facultad	Promedio-semestre	Promedio-acumulado	forma de pago-matricula		
					Presencial-virtual				Credito	Contado	Beca

Figura 1. Datos estudiante cada programa-cada facultad.

Datos demograficos estudiante						
Estrato	Genero		Edad	Ciudad	Estado civil		causa deserción
	M	F			S	C	

Figura 2. Datos estudiante cada programa-cada facultad.

En total son tres tablas con todos los registros discriminados en la imagen anterior de los periodos comprendidos entre el año 2002 hasta el año 2015, obteniendo así una base de datos no estructurada con un tamaño de 12,88 Gigas.

A continuación, se explica el contenido de las tablas creadas para la entrega de la base de datos. (Ver Tabla 3).

Tabla 3. *Datos requeridos para la investigación*

Datos requeridos para la investigación	
Código	El código del estudiante
Nombre	El nombre del estudiante
Apellido	El apellido del estudiante
Programa	El programa al cual pertenece o perteneció
Fecha inicio estudio	La fecha de inicio de estudio de un estudiante
Fecha fin estudio	La fecha de fin de estudio de un estudiante

Modalidad estudio	La modalidad de estudio de un estudiante (Presencial o a distancia)
Facultad	La facultad a la cual pertenece un estudiante
Estrato	El estrato del estudiante
Género	El género del o la estudiante
Edad	La edad del estudiante
Ciudad	La ciudad en la cual nació el estudiante
Estado civil	El estado civil del estudiante (Casado, soltero, unión libre, viudo, divorciado)
Tipo pago	El tipo de pago que ha hecho el estudiante (Crédito o contado)
Promedio acumulado	El promedio acumulado del estudiante.
Promedio acumulado ponderado	El promedio acumulado ponderado del estudiante.
Semestre	El semestre cursante o cursado del estudiante
Número matricula	El número de la matrícula del estudiante
Código	El código del estudiante

Con estos datos, se elaboró inicialmente una ficha técnica para caracterizar los datos entregados por la oficina de Informática de la Universidad Mariana – Tecno Smart.

A continuación, se presenta un fragmento de la ficha técnica, donde se realizó la caracterización de la base de datos. (Ver Figura 3).

BIGPRO

SISTEMA DE ANÁLISIS DE DATOS DE LA DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD MARIANA PASTO BASADO EN TÉCNICAS BIG DATA

FICHA TECNICA CARACTERISTICAS DE LOS DATOS

Estudiantes desertores de la Universidad Mariana Pasto.

TABLA	TIPO DE DATO	VOLUMEN	VARIEDAD	VERACIDAD	VELOCIDAD	UBICACIÓN DEL DATO	COMPLETO?
Promedio_semestre	Double	2.93 Gigas(48201 registros)	Datos numéricos	No campos nulos	200 ms	Oficina Sistemas TecnoSmart	Si
Estudiantes_por_programa	String	4,09 Gigas (59672 registros)	Datos String	No campos nulos	400 ms	Oficina Sistemas TecnoSmart	Si
Promedio_ponderado	Double	5,86 Gigas (225845 registros)	Datos numéricos	No campos nulos	500 ms	Oficina Sistemas TecnoSmart	Si

Figura 3. Ficha técnica para la caracterización de los datos.

De acuerdo a la base de datos entregada por la oficina Tecno-Smart, se ha elaborado una ficha técnica (Figura 3) para la caracterización de los datos en cuanto a tipo de dato, volumen, variedad, veracidad, velocidad, ubicación del archivo, la integridad del dato (si está completo o no) y una columna donde se especifica el nombre de cada tabla de la base de datos.

A continuación, se realiza una explicación del contenido de la ficha técnica para la caracterización de los datos de los estudiantes desertores de la Universidad Mariana.

Nombre tabla: en este ítem se especificó el nombre de la columna de la base de datos, por ejemplo: .promedio ponderado, semestre, edad, género, etc.

Para esta característica se identificó el nombre de la columna con respecto a la base de datos, en donde simplemente se evidencia el nombre de la columna con una longitud aproximada de 25 caracteres por columna, además todos los nombres se encuentran en mayúsculas.

A continuación, se muestra un ejemplo de algunas columnas seleccionadas para dar una idea de cómo se estructura estas columnas.

A6 ▾ ⋮ ✕ ✓ *fx* 1085268384

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	FECHA INICIO	FECHA FIN	MODALIDA	FACULTDAD	ESTRATO	GENERO	EDAD	CIUDAD	ESTADO CIVIL	TIPO DE PAGO				
2	30/01/13	08/06/13	Presencial	FACULTAD DE HUMANIDADES Y CIENCIAS SOCIALES	1	Femenino	26	SAN FRANCISCO	Soltero	CREDITO				
3	30/01/13	08/06/13	Presencial	FACULTAD CIENCIAS DE LA SALUD	2	Femenino	23	PASTO	Soltero	CREDITO				
4	30/01/13	08/06/13	Presencial	FACULTAD CIENCIAS CONTABLES, ECONÓMICAS Y ADMINISTRATIVAS	3	Femenino	25	PASTO	Soltero	CONTADO				
5	30/01/13	08/06/13	Distancia	FACULTAD CIENCIAS DE LA SALUD	1	Masculino	24	PASTO	Soltero	CONTADO				

Figura 4. Fragmento del contenido de la base de datos no estructurado.

En la Figura 4 se puede observar que en la fila 1 se encuentran los nombres de las columnas: fecha inicio, ficha fin, facultad, estrato, género, edad, ciudad, estado civil, tipo de pago, las cuales representan el tipo de dato que se encuentra en cada columna. En ocasiones estos nombres de columnas se llaman identificadores de datos o título de la columna. En total la tabla Excel tiene 13 columnas.

A continuación, se explica el tipo de dato que contiene la base de datos de la Universidad Mariana con respecto a los estudiantes desertores.

Tipo de dato: en este ítem se verificó el tipo de dato, en el cual se encontró datos de tipo Double, String, Int y Date.

Para ser más precisa la caracterización de los datos, se muestra con un fragmento de la base de datos (ver Figura 5), donde se identificó el tipo de dato que se puede encontrar. Por cuestiones de protección de datos personales, no se puede mostrar datos de los estudiantes de la Universidad Mariana, pero sí otros datos que no comprometen la privacidad y protección de los mismos, ya que solo se muestra información netamente investigativa y que no referencia a ningún estudiante.

ESTRATO	GENERO	EDAD	CIUDAD	ESTADO CIVIL	TIPO DE PAGO
1	Femenino	26	SAN FRANCISCO	Soltero	CREDITO
2	Femenino	23	PASTO	Soltero	CREDITO
3	Femenino	25	PASTO	Soltero	CONTADO
1	Masculino	24	PASTO	Soltero	CONTADO

Figura 5. Tipo de dato que se encuentra en la base de datos de la Universidad Mariana.

En la columna: estrato (ver Figura 3) se puede identificar que el tipo de dato es un valor numérico entero (Int), con una longitud máxima de una unidad, ya que en Colombia los estratos socioeconómicos en los que se pueden clasificar las viviendas y/o los predios son 6, denominados así:

1. Bajo-bajo.
2. Bajo.
3. Medio-bajo.
4. Medio.
5. Medio-alto.
6. Alto.

Por lo tanto, la principal característica de este dato es su longitud de una unidad (1 Und) y el tipo de dato entero; además se pudo identificar que los estratos que más se repiten son: 1, 2, 6. También se identificó que no se tiene registro de estrato de algunos estudiantes, por ello se depuró para mayor proximidad en la caracterización de la base de datos.

En la columna: edad (ver Figura 5) se pudo identificar que el tipo de dato es numérico entero (Int) con una longitud máxima de 2 unidades, ya que la edad promedio de un estudiante oscila entre 16 y 50 años de edad, utilizando así solo dos espacios en la celda correspondiente.

En la columna: ciudad (ver Figura 5) se pudo identificar que el tipo de dato es cadena de carácter (String) con una longitud máxima de 24 caracteres. La veracidad del dato de esta columna es aceptable, ya que escogió algunos registros para verificar el lugar de procedencia y así inferir en su veracidad e integridad del dato; además se identificó que este dato viene escrito en letra mayúscula.

En la columna: estado civil (ver Figura 5) se identificó que el tipo de dato es cadena de carácter (String) con una longitud máxima de 7 caracteres; en esta columna se depuró algunos registros, ya que existían algunos campos vacíos o nulos. Este dato cambia muy poco, ya que los estados que este contiene son similares en cuanto a longitud, integridad y volumen.

En la columna: crédito (ver Figura 5) se identificó que el tipo de dato es cadena de carácter (String) con una longitud máxima de 7 caracteres y su contenido está en mayúsculas y su variabilidad es muy poca, ya que solo tiene dos estados “crédito” y “contado”.

A continuación, se muestra y se explica el contenido de otro tipo de dato que se encontró en la base de datos (ver Figura 6).

B	C
PROMEDIO PONDERADO ACUMULADO	SEMESTRE
0	1
0	1
0	1
0	1
0	1
0	1
3,73	2
1,3	8
2,88	2
2,31	1
3,96	1
4,06	2

Figura 6. Fragmento de la base de datos obtenida mediante la oficina Tecno Smart.

En la columna: Promedio Ponderado Acumulado (ver Figura 6), se pudo identificar que el tipo de dato que se encuentra en las celdas es decimal (Double) entre el rango que va desde 0 hasta 5.0, esta es la calificación que maneja la Universidad Mariana. El dato tiene como longitud 3 espacios en la celda, el promedio ponderado acumulado se compone de una unidad, seguida del signo coma (,) con decenas y centenas. Además, el dato es variado en cuanto al rendimiento de cada estudiante.

En la columna: semestre (ver Figura 6) se pudo identificar que el tipo de dato es numérico entero (Int), utilizando solo un espacio en la celda, oscilando entre 1 y 10, los cuales representan los semestres que ha cursado un estudiante.

Estos son algunos de los datos que se han caracterizado; en su mayoría los datos tienen el mismo comportamiento o características, aunque es preciso decir que, existen otros datos que tiene otras características, las cuales se mencionarán más adelante.

Se realizó la caracterización en cuanto al tipo de dato que se encontró en la base de datos de la Universidad Mariana.

A continuación, se realiza una descripción de los datos en cuanto a volumen del mismo (ver Figura 7).

Volumen: en este ítem se mide el tamaño de los datos, pueden estar en el tamaño de las Gigas, Terabytes o Exabytes.

Para tener una idea precisa sobre el volumen de los datos, se muestra los últimos datos de una columna seleccionada (ver Figura 7), para identificar cuántos datos puede tener una tabla en la base de datos.

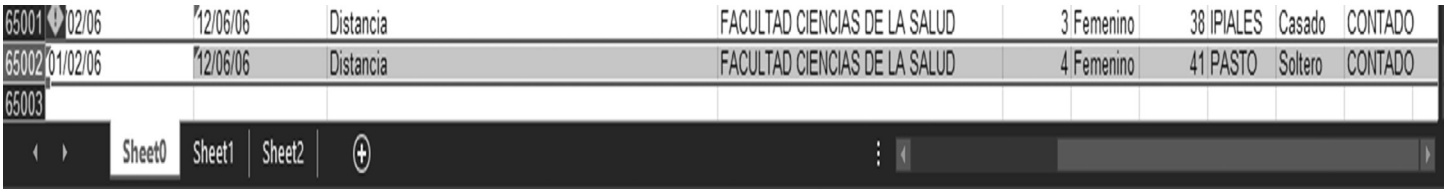

Figura 7. Fragmento de la base de datos ofrecida por la oficina Tecno Smart.

El último registro se puede identificar en la fila 65003 (ver Figura 7), representa que en la hoja actual llamada Sheet0 se encuentran 65.003 registros, en la hoja Sheet1 se encuentra otro registro con la misma cantidad de registros y en la tercera hoja Sheet2 se encuentran 59.673 registros, aplicando matemáticas simples como la suma, se obtiene en total 189.679 registros, con un volumen de 2.5 Gigas.

A continuación se explica la velocidad de los datos, en cuanto a procesamiento, fueron expresados en milisegundos.

Velocidad: en este ítem se verificó la velocidad en que el dato se demora en ejecutarse (ver Figura 8), entiéndase por tiempo de ejecución, el momento en que se abre el archivo para ser procesado y éste depende del volumen del dato. Para la base de datos recolectado los tiempos estimados fueron los siguientes:

Tiempo de ejecución de cada tabla de la base de datos (Velocidad)	
Nombre representativo tablas de la base de datos	**Tiempo de ejecución (ms)**
Estudiante por programa	200 ms
Promedio ponderado por semestre cada estudiante	400 ms
Promedio acumulado por semestre cada estudiante	500 ms

Figura 8. Tiempo de ejecución de cada tabla de la base de datos (velocidad).

Con lo anterior, se puede deducir que los datos son muy variados y se encuentran datos numéricos, datos de tipo fecha, datos de cadena de caracteres, datos de tipo decimal como por ejemplo: promedio de un estudiante; además, la veracidad de los datos es confiable, ya que no existen campos nulos en las tablas correspondientes en la base de datos, en cuanto a volumen, los datos en esencia son rápidos a la hora de ejecutarlos. Para esta investigación, la base de datos es aceptada para su posterior tratamiento y análisis, ya que en la caracterización de la base de datos se evidenció un índice de aceptación del 89 %, de acuerdo con los criterios de caracterización establecidos en la ficha técnica.

Referencias

Hernández, R., Fernández, C. y Baptista, P. (2003.). *Metodologías de la investigación.* . México, D.F: McGraw-Hill Interamericana.

Jürgen, H. (1977). *Nuevo enfoque de la filosofía transcendental.* 7(3/4), 347-352.

Ministerio de Educación. (2009). *www.mineduacion.gov.co*. Recuperado de www.mineduacion.gov.co

Ministerio de Educación Superior Colombiano. (2008). Recuperado de www.mineducacion.gov.co/spadies/

Publicaciones Semana S.A. (2015). 5000 Empresas. *Revista Dinero*. Recuperado de http://www.dinero.com/edicion-impresa/caratula/articulo/articulo-apertura-5000-empresas-mas-grandes-del-pais-segun-revista-dinero/209392

Secreteria de educacion de Nariño. (2014). *www.sednarino.gov.co.* Recuperado de www.sednarino.gov.co

Universidad EAFIT. (1999). Deserción estudiantil universitaria en Colombia. *Revista universitaria EAFIT.*

Universidad Mariana. (1970). Recuperado de *www.umariana.edu.co/docinstituciones/mision_vision_directrices_institucionales.pdf*.

Universidad Nacional a Distancia. (2014). Recuperado de https://www.unad.edu.co/