Revista Colombiana de Computación, Vol. 20, No 2 julio-diciembre 2019, pp. 6-19
http://dx.doi.org/10.29375/25392115.3717

Aplicación para la gestión y el análisis de información relacionada con la deserción estudiantil universitaria

Application for the management and analysis of information related to the university student desertion

Luz Yamile Caicedo Chacón[1], Cristian Noé Cárdenas Parra[1], Juan Sebastián Müller Rueda[1], Jeniffer Tatiana Ortiz Bernal[1]

[1]Facultad de Ciencias Naturales e Ingeniería, Fundación Universitaria de San Gil – UNISANGIL, San Gil, Colombia
lcaicedo@unisangil.edu.co, cristiancardenasp@unisangil.edu.co, juanmuller@unisangil.edu.co, jenifferortiz@unisangil.edu.co

(Recibido: 26 marzo 2019; aceptado: 8 julio 2019)

Resumen

Una de las principales metas del Ministerio de Educación Nacional (MEN) es aumentar el índice de permanencia en las Instituciones de Educación Superior (IES). En consecuencia, se han definido estrategias que se espera que permitan alcanzar la graduación y garantizar la permanencia de los estudiantes que se encuentran cursando un programa técnico, tecnólogo o profesional. Por consiguiente, la Fundación Universitaria de San Gil – UNISANGIL se propuso desarrollar e implementar una herramienta software bajo la metodología SCRUM, la cual permitirá manejar los registros de la información obtenidos a partir de los datos personales, socioeconómicos, psicosociales y los derivados del desempeño académico de los estudiantes antes y durante su permanencia en la institución, con el fin de generar estadísticas e informes (periódicos y específicos) utilizando técnicas de minería de datos y software de inteligencia de negocios para proporcionar una herramienta que apoye el proceso de toma de decisiones y la búsqueda de patrones para la detección de riesgos sobre la deserción.

Palabras clave: Inteligencia de negocios, Alertas tempranas, Minería de datos, Deserción estudiantil, Weka.

Abstract

One of the main goals of the Ministry of National Education (MEN) is to increase the permanence index in higher education institutions, IES. Consequently, strategies have been defined to achieve graduation and ensure the permanence of students who are studying a program, technical, technological or professional. Therefore, the University Foundation of San Gil - UNISANGIL was proposed to develop and implement a software tool under the SCRUM methodology, which will manage the records of information obtained from personal, socioeconomic, psychosocial and academic performance data of the students before and during their stay in the institution, in order to generate statistics and reports, periodic and specific; using data mining techniques and business intelligence software to provide a tool that supports the decision-making process and the search for patterns for the detection of risks on desertion.

Keywords: Business intelligence, Early alerts, Data mining, Student desertion, Weka.

Cite this work as Caicedo Chacón L., Cárdenas Parra C., Müller Rueda J., Ortiz Bernal J. (2019). Aplicación para la gestión y el análisis de información relacionada con la deserción estudiantil universitaria Revista Colombiana de Computación, 20(2), 6-19. https://doi.org/10.29375/25392115.3717

1. Introducción

En Colombia la deserción estudiantil tiene un gran impacto que genera ciertas probabilidades de baja de productividad laboral por falta de personal capacitado. Esto obstaculiza los avances sociales hacia los objetivos del Gobierno y desequilibra financieramente a las universidades (Riveros Garzón, 2016). Aunque se ha ido posicionando la permanencia y la finalización de los estudios como eje central en las agendas de las instituciones y en las políticas nacionales de los últimos años, no es suficiente. Se debe continuar trabajando en el sector para disminuir la deserción anual que actualmente se encuentra en un 8% a nivel universitario y en 15% a nivel técnico profesional y tecnólogo, de acuerdo con el Plan Nacional de Desarrollo 2014-2018 "Todos por un nuevo país" (MinEducación, 2015).

La evolución tecnológica ha generado nuevos campos de investigación que de forma permanente producen adelantos que pretenden facilitar la vida del ser humano, como el acceso a grandes cantidades de información, el desarrollo de tareas en tiempos reducidos y la capacidad de contar con reportes en tiempo real, entre otros. En consecuencia, se han construido aplicaciones software cada vez más robustas en distintas áreas del conocimiento (Ballesteros Román, Sánchez-Guzmán y García Salcedo, 2013).

La trascendencia de las Tecnologías de la Información y las Comunicaciones (TIC) y el aporte de la inteligencia de datos con respecto a la sociedad y, en especial, en el sector educativo juegan un papel sumamente importante, ya que han influido en la teoría de la pedagogía y el aprendizaje. Además, han impulsado la mejora del software educativo, sobre todo con respecto a su capacidad para personalizar la experiencia del estudiante.

Uno de los grandes retos de las universidades es la necesidad de obtener datos para identificar las causas de deserción de los estudiantes en el transcurso de su carrera. La información es indispensable para disminuir este índice. Así mismo hay que identificar con anticipación las situaciones que llevan a un estudiante a abandonar su proceso de formación en cualquier momento, haciendo monitoreo para poder tomar acciones correctivas. Una de las herramientas utilizadas en la educación que permite identificar información potencialmente útil en grandes cantidades es la Minería de Datos Educativa (MDE), que llega a ser comparada con los diferentes paradigmas más tradicionales de investigación referente a la educación y que tiene por objetivo generar labores como clasificación, agrupamiento (*clustering*), mejoramiento de los sistemas de aprendizaje y enseñanza y predicción de comportamientos a futuro, entre otros. Esto ofrece una variedad de ventajas que se centran en el desarrollo de métodos de descubrimientos, gracias a los datos de las plataformas educacionales para tener un mejor entorno de comprensión y aprendizaje (Jiménez Galindo y Álvarez García, 2010).

Además, las Instituciones de Educación Superior (IES) tienen sistemas de información que registran datos personales, socioeconómicos y sobre el desempeño académico de los estudiantes durante su proceso de formación (Ballesteros Román et al., 2013). El MinEducación cuenta con el Sistema para la Prevención de la Deserción de la Educación Superior (SPADIES), que obtiene datos de las universidades del país y sirve como referente para generar estrategias de promoción de la permanencia y graduación estudiantil. Por medio de este software se monitorean y analizan las variables que conforman los factores determinantes de la deserción (MinEducación, 2015).

Sin embargo, el SPADIES aun siendo una herramienta que ayuda en esta problemática, no es suficiente para bajar los niveles de deserción en las IES. Es por esto, que se recomienda a cada una de ellas realizar el tratamiento de la información particular para hacer análisis propios sobre el fenómeno de la deserción. La Fundación Universitaria de San Gil cuenta con diferentes herramientas software para la gestión de la información, pero estos no poseen las utilidades que se requieren para hacer seguimiento y acompañamiento a los estudiantes, por lo que se ve la necesidad de crear un software transaccional y analítico que genere reportes y alertas para una mayor efectividad en el proceso de acompañamiento al estudiante.

Se tomaron como base los lineamientos que MinEducación (2015) definió para abordar el problema de la deserción y que hacen referencia a cuatro factores con sus respectivas variables. Estos son:

1. Individuales: edad, sexo, estado civil, posición dentro de los hermanos, entorno familiar, posibles calamidades, problemas de salud, integración social, incompatibilidad horaria con actividades extraacadémicas, expectativas satisfechas o embarazo.
2. Académicas: orientación socio ocupacional, tipo de colegio, rendimiento académico, calidad del programa, métodos de estudio y aprendizaje, pruebas Saber, resultados en el examen de ingreso, cualificación docente y grado de satisfacción con el programa.

3. Institucionales: normalidad académica, servicios de financiamiento, recursos universitarios, orden público, entorno político, nivel de interacción entre estudiantes y docentes, apoyo académico y apoyo psicológico.
4. Socioeconómicas: estrato, situación laboral del estudiante, así como de los padres o acudientes, calidad de dependencia económica o personas a cargo, nivel educativo de los padres y entorno macroeconómico del país.

De acuerdo con estos factores, UNISANGIL, en el marco de los estudios institucionales de pertinencia, grupo permanencia y deserción, diseñó un instrumento denominado 'Encuesta de deserción', la cual fue aplicada a todos los programas en las tres sedes. Para la realización del primer piloto con el software, se trabajó con el Programa de Enfermería y los programas de la Facultad de Ingeniería. Estos datos fueron recolectados y analizados inicialmente con estadística descriptiva utilizando el software SPSS y, posteriormente, se procesaron con la herramienta de minería de datos WEKA. Se utilizaron dos técnicas que WEKA proporciona una para el análisis sobre las cuales se basa en el algoritmo *K-Means* y, en segundo lugar, la clasificación soportada en la técnica conocida como J48.

La herramienta (transaccional y de inteligencia de negocios) fue diseñada para ser utilizada por todos los integrantes de la comunidad universitaria de UNISANGIL. Las pruebas iniciales de los formularios de caracterización se realizaron con estudiantes del programa de Ingeniería de Sistemas y el despliegue del caso de prueba del software se realizó con el Programa Enfermería de la Sede San Gil. Actualmente, los demás programas de las tres sedes de la institución (San Gil, Yopal y Chiquinquirá) están usando la aplicación para realizar la caracterización de los estudiantes que ingresan a primer nivel de aprendizaje. El Programa de Apoyo y Seguimiento Académico (PASA) realiza la generación de los reportes que en cada semestre académico se deben generar para presentar informes ante los respectivos comités, además de cumplir con las labores de acompañamiento a estudiantes.

2. Metodología de desarrollo del software

Para la construcción del software se optó por la metodología de desarrollo ágil SCRUM, debido a que está compuesta por buenas prácticas para trabajar en equipo y obtener los mejores resultados en un proyecto. Además de permitir incluir los cambios que se presentan en las directrices del MinEducación para abordar la deserción, esta metodología es útil para adicionar nuevos requerimientos sobre el proceso de desarrollo (Cárdenas Parra, Müller Rueda y Ortiz Bernal, 2019).

Según Trigas (2012), las metodologías ágiles surgen como una alternativa a las tradicionales, las cuales son demasiado pesadas para las actuales condiciones del mercado. El desarrollo ágil está centrado en la comunicación, la iteración y en reducir elementos intermedios. Así mismo, permite la participación del cliente de forma activa para validar la interpretación que hace el equipo de desarrollo de las historias de usuario y de esta manera garantizar un buen modelamiento del sistema antes de empezar a codificar, logrando, a través de la comunicación, la resolución de diferencias al interior del equipo, facilitando la toma de decisiones.

Además, SCRUM trabaja sobre un marco iterativo e incremental para la construcción de software. Se apoya en los *sprints*, que son los ciclos de trabajo que definen las iteraciones con espacio de entre 1 a 4 semanas y se desarrollan consecutivamente (Mariño y Alfonzo, 2014). Para dar inicio a un *sprint*, el equipo de trabajo elige los requisitos del cliente que se van a abordar a partir de una lista priorizada, con el compromiso de terminar el desarrollo de estos al final de la tarde. Durante la ejecución de la fase no se pueden cambiar los requisitos. Al final del *sprint,* el equipo presenta a los participantes del proyecto lo que ha construido para ser revisado y aprobado (Figueroa, Solis, y Cabrera, 2008). Y así poder ir conceptualizando y poder transformar del lenguaje natural a código de programación las necesidades que el software tenderá a resolver.

Una vez terminada la recolección de información, se continuó con la siguiente etapa: desarrollo por módulos mediante los *sprint.* Primero, la construcción de los casos utilizando la herramienta StarUML, los diagramas de clases, los diagramas de secuencias, el modelo de bases de datos y el diseño de interfaz gráfica de usuario, todo para facilitar la comprensión y el entendimiento del funcionamiento del sistema que va a ser automatizado. Se siguió con el proceso del software hasta llegar a la integración del usuario desarrollado, se validó la funcionalidad para presentar los avances y, una vez aceptado el módulo, se pudo pasar a uno nuevo. Así, se creó incrementalmente el software, dependiendo del alcance de la aplicación y las necesidades que tenía el Programa de Apoyo y Seguimiento Académico (PASA) de UNISANGIL.

Una vez se finalizó el *sprint,* se revisaron las prioridades del PASA para dar continuación al proyecto con el desarrollo de un nuevo *sprint*. De esta manera se obtuvo un producto a nivel transaccional que está listo para ser mostrado a los usuarios finales: docentes, estudiantes y funcionarios del PASA. El diseño del esquema metodológico del software se presenta en la Figura 1:

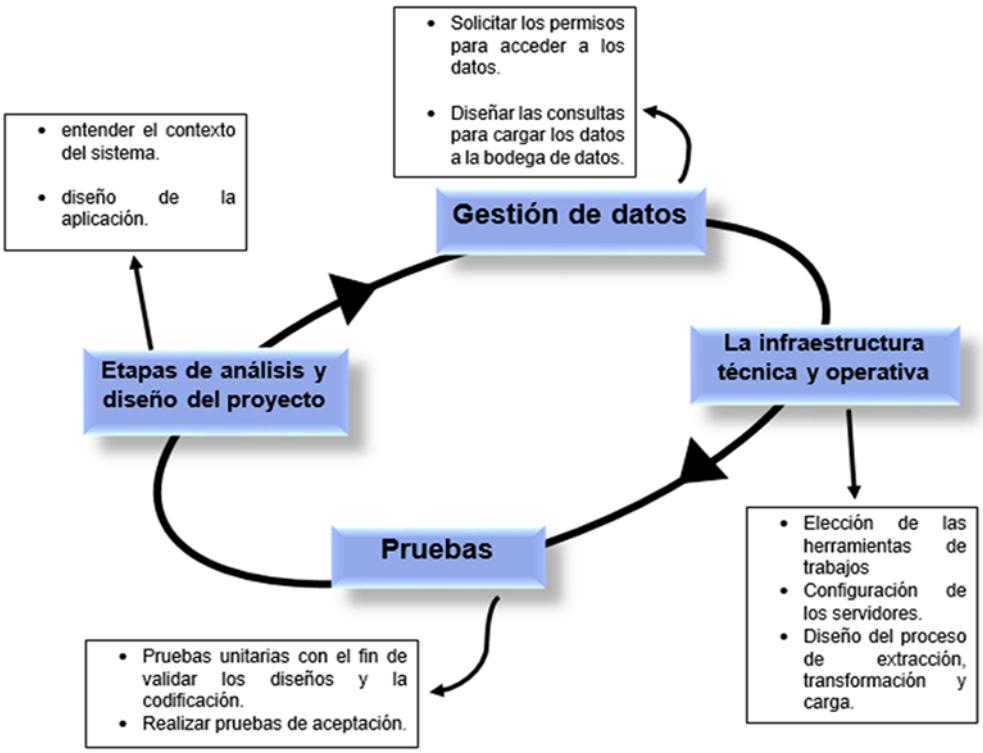

Figura 1. Esquema metodológico

Una vez se tuvo el software a nivel transaccional se inició la fase que permitió tomar la información de la caracterización y el apoyo en diferentes fuentes de datos como los estudios institucionales de pertinencia, en especial, el estudio relacionado con la deserción para tomar como referente el instrumento que se desarrolló y que sirviera como fuente de datos para poder analizarlos a través del software de minería de datos WEKA.

3. Metodología para el análisis de datos

Para el análisis de datos se trabajó con la metodología CRISP-DM. Según Galán (2015), la metodología propuesta por el proceso estándar de la industria para la minería de datos, CRISP-DM es un modelo que describe la manera en la que los expertos en esta materia abordan el problema. El proceso está organizado en seis fases para el ciclo de vida de un proyecto, cada una de ellas a su vez estructuradas en varias tareas generales. En el segundo nivel las tareas generales arrojan actividades específicas, donde se detallan los trabajos que deben ser desarrollados para situaciones determinadas.

La primera fase se encarga del análisis del problema, que permite conocer los objetivos y requerimientos del proyecto, dando una perspectiva empresarial, con el fin de transformarlos en objetivos técnicos y en un plan de trabajo. En la segunda fase se realiza el análisis de datos, para poder construir una idea del problema, además de describir la calidad de los datos y poder establecer una hipótesis inicial (Galán Cortina, 2015; Rodriguez Montequín, Álvarez cabal, Mesa Fernández y González Valdés, 2005).

Las siguientes dos fases preparación de datos y modelado trabajan de forma sistemática, ya que interactúan dependiendo de las técnicas de modelado junto con los criterios más adecuados para un proyecto de minería de datos. Por consiguiente, los datos que se van a utilizar necesitan ser procesados de diferentes maneras.

En la fase de evaluación es necesario revisar los datos obtenidos hasta el momento, por si se presentan errores, y continuar con la evaluación del modelo, considerando la ejecución de los criterios de resolución del problema. Generalmente en la fase de explotación, los proyectos de minería de datos continúan con la documentación, la entrega de los resultados de manera evidente, la constante revisión de la herramienta y la comunicación de los resultados.

En la Figura 2 se pueden observar las fases en las que se divide CRISP-DM y las posibles secuencias a seguir entre ellas.

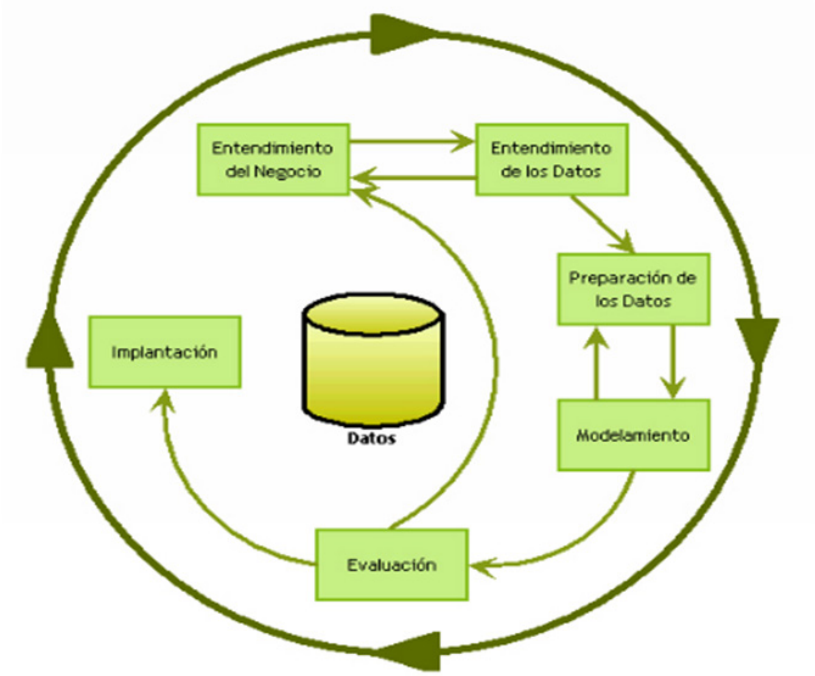

Figura 2. Secuencia del proceso CRISP-DM.
Fuente: Tomado de Galán Cortina, 2015.

4. Metodología del diseño arquitectónico

En UNISANGIL se utiliza CentOS como sistema operativo de servidores. Además, el software utilizado para el alojamiento del servidor web es LAMPP, el cual corre sobre ambiente Linux. La herramienta construida se denomina Sistema de Alertas Tempranas Académicas de UNISANGIL (*Academic Early Warning System* – AEWS) y ya se encuentra alojado en los servidores de la Universidad.

El software transaccional se alimenta de la información que proviene de AcademuSoft, el cual está relacionado con los horarios de clase, asignaturas, matrículas que realizan los estudiantes y aulas de clase para todos los semestres, incluyendo también la información de los docentes. Se cuenta con información registrada en hojas electrónicas que contienen los resultados de las pruebas genéricas de conocimiento debido a que estas se desarrollan en otra plataforma y es en formato Excel, donde se descarga la información que se tiene para generar los reportes. Otras fuentes, son SPADIES y Sistema Nacional de Información de Educación Superior (SNIES), que son sistemas a los cuales la Universidad reporta información y sirve de insumo para el Sistema de Alertas Tempranas Académicas de UNISANGIL.

El desarrollo del software se trabajó en varios módulos que son Estudiante, Docente, Administrador y Psicólogo. El módulo Psicólogo permite registrar las entrevistas que realiza el profesional de admisiones a los aspirantes. El sistema genera una alerta que se envía a través de correo electrónico al psicólogo y al funcionario de admisiones informándoles que la entrevista ha sido realizada y comunicándoles si se genera una alerta acerca de ese estudiante el cual debe ser remitido a otra instancia de atención.

La herramienta que se utilizó para el envío de correos electrónicos automáticos fue PHPMailer a través del plugin de Google, que es el sistema de correos que maneja la universidad. En inteligencia de negocios se usó el software Pentaho versión 8.0 que corresponde a una *suite* desarrollaba bajo licencia de software libre. Esta tiene comunicación con diferentes motores de base de datos, permitiendo alimentar las estructuras de almacenamiento de información: bases de datos y bodegas de datos.

La bodega de datos, junto con la base de datos que se construyó, está alojadas en servidores Oracle como servidor de gestor de base de datos y Pentaho se comunica con este motor para poder realizar el proceso de extracción, transformación y carga que comprende el paso de los datos de la base de datos hacia la bodega de datos. Teniendo los datos en la bodega, se pueden construir los cubos de información para generar los informes y alertas a partir de la información que no se trabaja en el sistema transaccional. Lo anterior se debe a que el sistema transaccional se usa únicamente para el periodo que se está cursando y las demás alertas son a nivel histórico. Estas se van a manejar en la bodega de datos.

Una representación de la Metodología del diseño arquitectónico se da a conocer en la Figura 3.

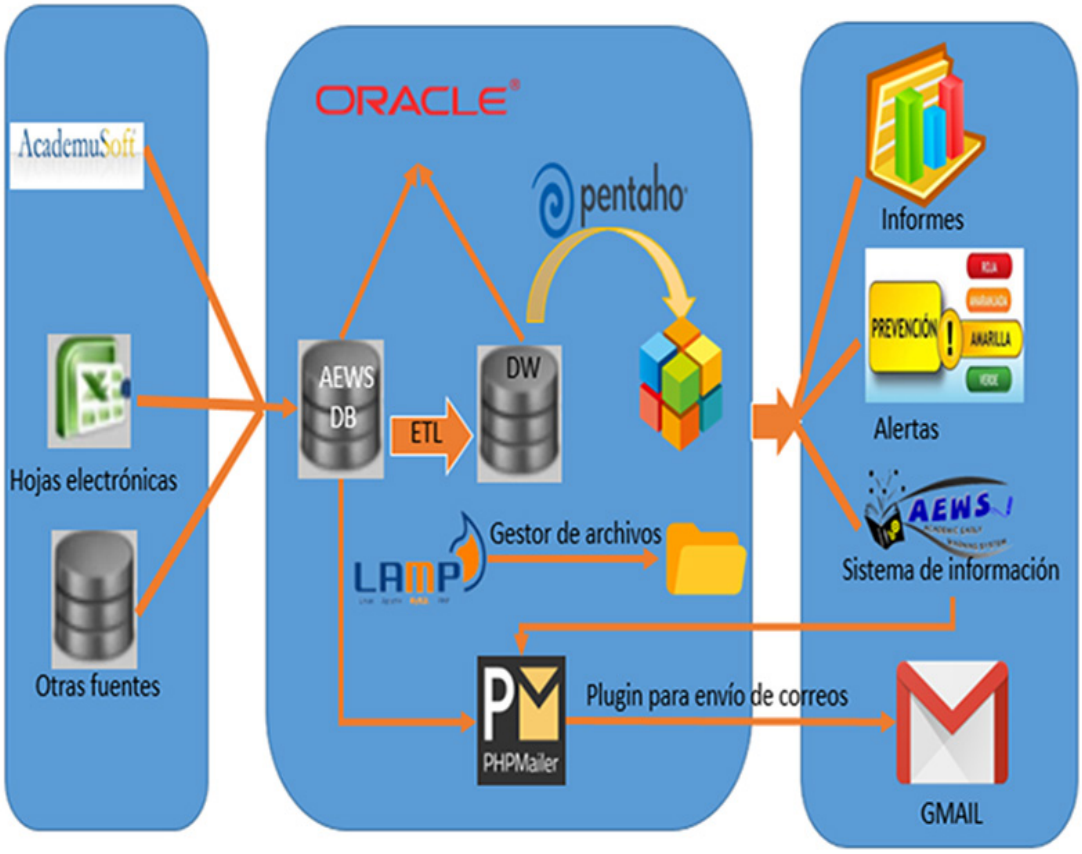

Figura 3. Metodología del diseño arquitectónico.

5. Implementación del software

Para la fase de implementación, UNISANGIL facilitó el alojamiento del software dentro de los servidores de la institución, otorgando permisos de acceso al equipo desarrollador. A través de internet, los demás usuarios como docentes, estudiantes y funcionarios de PASA tienen acceso al sistema, lo que conlleva que podría ser utilizado desde cualquier lugar, incluyendo las diferentes sedes de la institución. A su vez, la herramienta fue diseñada para ser usada desde cualquier dispositivo móvil.

Durante la construcción de la Interfaz Gráfica de Usuario se contó con la participación de los diferentes grupos de usuarios a través del desarrollo de reuniones con algunos representantes de estos,

donde se mostraba el diseño y el correcto funcionamiento de los diferentes módulos desarrollados, ítems que fueron aprobados por aquellos mismos usuarios.

Para realizar la capacitación sobre el uso del software, se realizaron reuniones con docentes y estudiantes, a fin de mostrar los avances en los módulos del software, y obtener los comentarios de los usuarios para validar el correcto registro de la información que se almacena en la base de datos y la aplicación de encuestas de satisfacción que permitieron conocer el concepto generado a partir del uso de la interfaz desarrollada.

Para los estudios que se realizarán posteriormente, se ha planteado que cada semestre académico, cada estudiante deba ingresar al sistema a registrar las modificaciones pertinentes, a fin de poder llevar un registro histórico, bajo el cual se aplicará el seguimiento minucioso que permita prever los riesgos de deserción.

6. Resultados

Para desarrollar el software transaccional fue preciso entender el funcionamiento de los procesos y tareas del PASA, además de incluir temas relacionados con la estadística y la psicología, junto con los lineamientos definidos por el MinEducación a través de la Guía para la implementación del modelo de gestión de permanencia y graduación estudiantil en instituciones de Educación Superior (2015). Por lo cual fue necesario definir la secuencia de actividades que el sistema debía realizar, considerando los flujos de información que maneja la Unidad de Permanencia y Graduación Estudiantil de UNISANGIL y la capacidad de respuesta del PASA, que condujo a la elaboración del diagrama de transición de estados que se ilustra en la Figura 4.

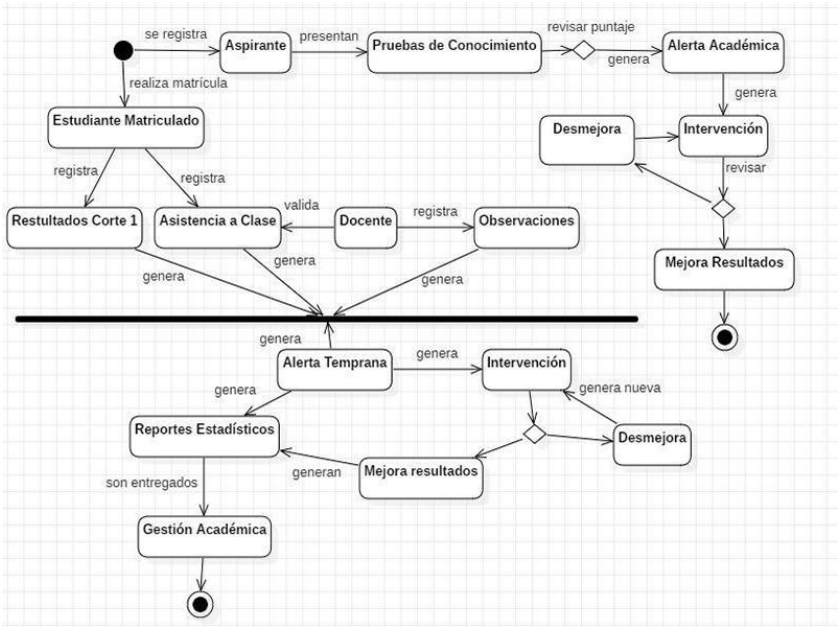

Figura 4. Diagrama de transición de estados AEWS.

A continuación, teniendo como referente el SNIES junto con la información que reporta semestralmente la universidad, se diseñó la base de datos junto con los demás formularios que componen la aplicación. Es decir, inicio de sesión para todos los usuarios del sistema, clasificación por roles, formulario de datos personales, familiares, académicos y financieros del estudiante, datos personales del docente, formulario de asistencia (Figura 5), entrevista del psicólogo, gestión de usuarios y reportes de datos estadísticos.

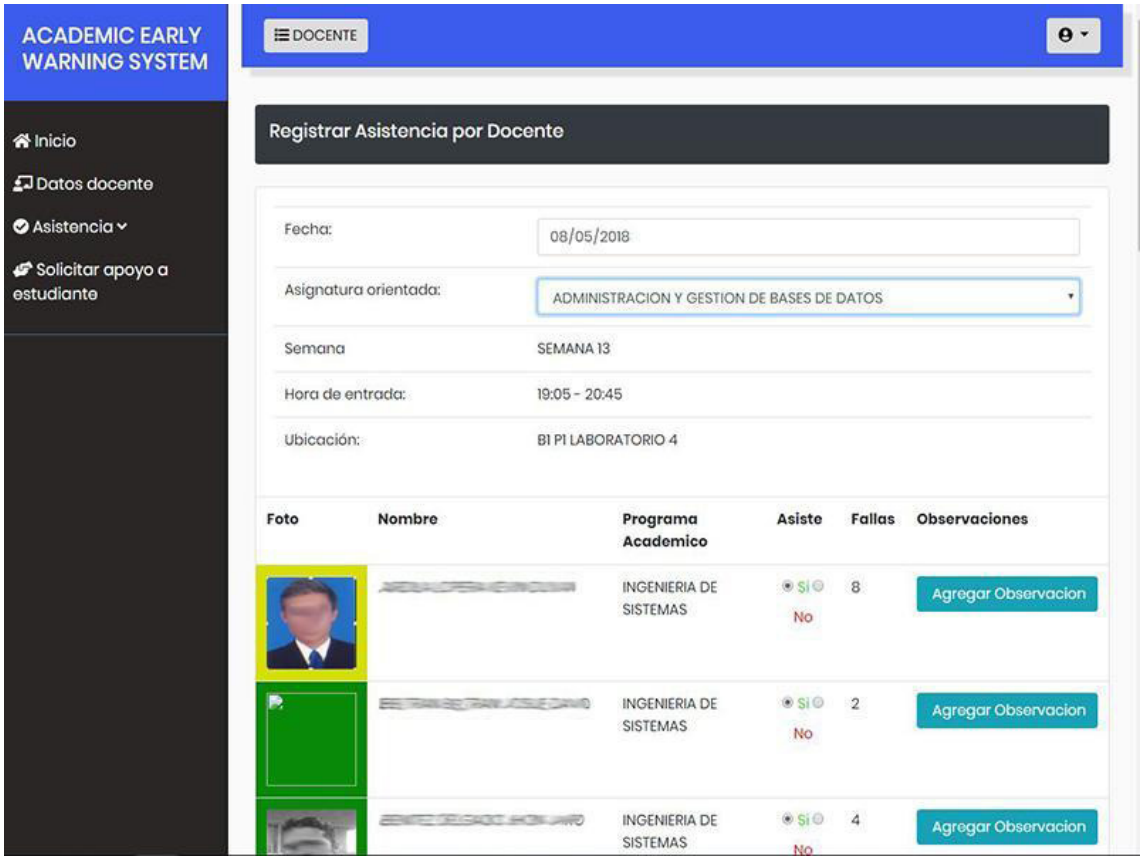

Figura 5. Registro de Asistencia.

Una vez construidas las interfaces para la caracterización de estudiantes, se aplicó el primer piloto con estudiantes del Programa de Ingeniería de Sistemas, para validar la aceptación, por parte de los estudiantes, de los formularios construidos. Posteriormente, se trabajó con el Programa de Enfermería para obtener el registro de la caracterización de los estudiantes de primero a último nivel de aprendizaje, matriculados durante el periodo 2018-02. La interfaz fue elaborada considerando los factores determinantes de la deserción planteados por el MinEducación. La Figura 6 presenta, como ejemplo, el primer formulario, que corresponde a datos personales del estudiante.

A continuación, se elaboraron las interfaces para el rol de Psicólogo, en el cual se tomaron las preguntas para realizar la entrevista a los estudiantes admitidos. Todo este rol mantiene las características solicitadas por el usuario (Psicólogo). En la Figura 7 se presenta la pantalla de inicio de dicho rol.

7. Técnicas predictivas

A continuación, se presentan los modelos más conocidos en el software WEKA, los cuales se trabajaron en esta investigación con el fin de determinar el modelo apropiado para proporcionar los patrones de deserción en UNISANGIL, específicamente en el Programa de Enfermería de la Sede San Gil. Se realizó la clasificación de los estudiantes mediante la utilización de la información de carácter socio económico, académico, institucional e individual.

7.1 Bayes Net

Según Santiesteban Rojas (2012), para evaluar estas probabilidades se han planteado numerosos algoritmos, entre los que cabe destacar el BayesNet, uno específicamente de enseñanza, práctico más manejados por su sencillez, con poco tiempo para el procesamiento y alto poder predictivo. Esta técnica es uno de los modelos de clasificación más efectivos. Están basadas en las redes bayesianas, que son modelos gráficos probabilísticos.

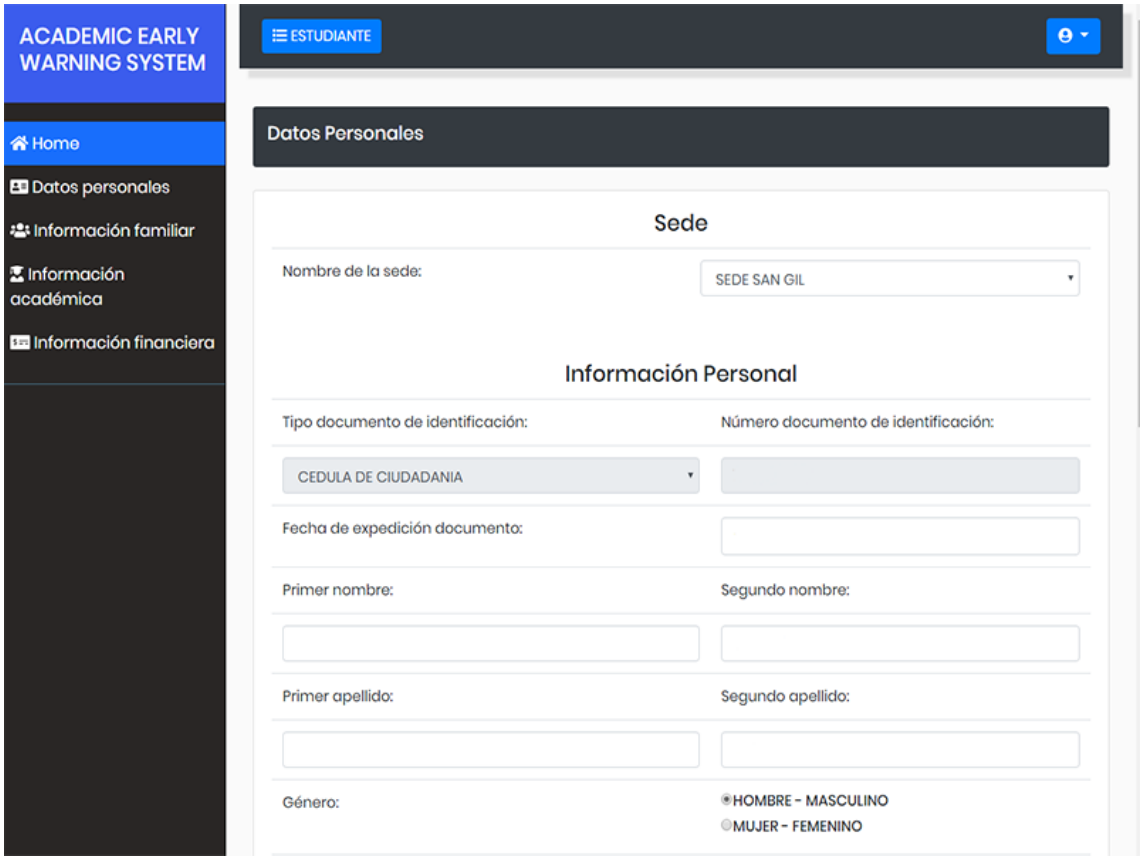

Figura 6. Datos personales del estudiante.

Figura 7. Inicio psicólogo.

El modelo utilizó como entrada la información obtenida de las encuestas de deserción, además la configuración de los experimentos se aplicó la validación cruzada con diez iteraciones para entrenar el modelo. La validación cruzada fracciona los datos encasillados en conjuntos de entrenamiento y de prueba (Santiesteban Rojas, 2012).

Para valorar la rentabilidad del modelo se utilizó el operador *Cross-Validation*. Este operador admite determinar la evolución de validación cruzada con *10-fold* sobre el conjunto de datos de entrada para evaluar el algoritmo de aprendizaje. El desempeño del modelo se evaluó con el operador *Simple Estimator*. Este operador presenta los resultados de desempeño del algoritmo en términos de exactitud, precisión,

recall, error y curva ROC. Para analizar los errores generados a partir de un modelo clasificación se emplea la matriz de confusión.

7.2 Árbol J48

Es una técnica usada para la clasificación de los datos. Este algoritmo genera un árbol de decisión de forma recursiva al examinar el criterio de la mayor proporción de ganancia de información, donde una instancia es clasificada siguiendo el camino de condiciones, desde la raíz hasta llegar a una hoja, la cual corresponderá a una clase etiquetada. Es decir, elige el atributo que mejor clasifica los datos. Se puede convertir fácilmente en un conjunto de reglas de clasificación (Vizcaíno Garzón, 2008).

Para la creación de este modelo se tuvo en cuenta la misma validación cruzada que se empleó en la técnica BayesNet y de las mismas características de desempeño del modelo anterior, utilizando *10-fold* sobre el conjunto de datos de entrada.

8. Técnicas descriptivas

Agrupamiento o *clustering*: Este algoritmo pertenece al aprendizaje no supervisado; que consiste en una variedad de elementos que pertenecen a diferentes o a un mismo grupo y que se basan en alguna característica concreta. Es decir, poder descubrir relaciones de manera implícita en un conjunto de datos que seguramente no contemplamos a simple vista o no están asignados previamente (Rodríguez, 2016).

K-means: Es una de las técnicas implementadas en Minería de datos. Este algoritmo es uno de los más utilizados para realizar agrupamiento. Además de su eficiencia, es sencillo de aprender e implementar. Su objetivo es ubicar todos los elementos en un lugar determinado y, dependiendo de sus características, formar grupos de datos con similitudes, pero diferentes a los demás que integran otros grupos (Lara Gutierrez, Lara Ruíz, Hernández Hernández, Hernández Hernández, y Hernández Hernández, 2016).

9. Análisis de datos

Después de tener el software en funcionamiento y realizando pilotajes, se continua con la fase de comprensión de los datos, en la que se utilizaron técnicas de visualización con el objetivo de realizar una exploración preliminar de los registros y validar la calidad de los datos. Por lo cual, esta primera tarea es obtener las fuentes de los datos del sistema de información académica de UNISANGIL en el caso de estudio del Programa de Enfermería. El primer conjunto de datos agrupa la información personal y familiar como son: edad de ingreso, género, lugar de procedencia, estrato socioeconómico, origen étnico, estado civil, SISBEN, situación económica del estudiante (Dependiente, Independiente, Empleado y Otro) y ocupación y nivel educativo de los padres.

Como segundo conjunto de datos se tomó la información del historial académico y financiero de los estudiantes tales como: carácter del colegio (académico, académico y técnico, o técnico), tipo de colegio (privado o público), metodología, puntaje obtenido por el estudiante en las pruebas Saber 11, quién costea los estudios (padres, becas, empresa, crédito educativo o el mismo estudiante), tiempo que tardó en ingresar a la institución después de graduado de bachiller, valor de la matrícula de la universidad, programa en el que se encuentra matriculado, apoyos recibidos por el ente universitario y si el estudiante trabaja mientras estudia.

Las consultas generadas se realizaron a través del sistema de gestión de base de datos Oracle. Se realizó la extracción de los datos en forma de concatenación, adquiriendo un archivo plano con 41 atributos y 174 registros de los estudiantes matriculados del Programa de Enfermería. Una vez ejecutado el análisis se procedió con la etapa de preparación de la información que contenía las tareas de selección de los datos a los que se le van a aplicar la técnica de modelado para su respectivo análisis.

El análisis exploratorio es una tarea que permite examinar en detalle algunas variables e identificar características. Para este, se emplearon algunas de las herramientas de visualización como las tablas y gráficos, con el propósito de describir los objetivos de minería de datos de la fase de comprensión.

A partir de los datos recolectados y analizados, se continuó con la selección del modelo descriptivo. El algoritmo escogido fue el de agrupamiento, dado que, en la verificación del estado del arte ejecutado, se estima el más conforme con los objetivos de minería de datos propuestos en esta investigación.

La técnica de árboles de decisión se empleó para la tarea de organización. Además, el modelo de clasificación que utiliza es posiblemente el más empleado por su simplicidad. También se escogió esta

técnica ya que, como se puede apreciar en las Figuras 8 y 9, el árbol J48 presenta un 96% de las instancias correctamente clasificadas, mientras que el modelo de BayesNet ofrece un 88%.

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          24               96      %
Incorrectly Classified Instances         1                4      %
Kappa statistic                          0.918
Mean absolute error                      0.0789
Root mean squared error                  0.2072
Relative absolute error                 15.8861 %
Root relative squared error             41.4501 %
Total Number of Instances               25

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,091    0,933      1,000    0,966      0,921    0,912     0,875     SI
               0,909    0,000    1,000      0,909    0,952      0,921    0,912     0,949     NO
Weighted Avg.  0,960    0,051    0,963      0,960    0,960      0,921    0,912     0,908

=== Confusion Matrix ===
```

Figura 8. Resultado árbol J48.

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          22               88      %
Incorrectly Classified Instances         3               12      %
Kappa statistic                          0.7492
Mean absolute error                      0.1742
Root mean squared error                  0.3413
Relative absolute error                 35.0765 %
Root relative squared error             68.2817 %
Total Number of Instances               25

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,273    0,824      1,000    0,903      0,774    0,890     0,872     SI
               0,727    0,000    1,000      0,727    0,842      0,774    0,890     0,918     NO
Weighted Avg.  0,880    0,153    0,901      0,880    0,876      0,774    0,890     0,892

=== Confusion Matrix ===

  a  b   <-- classified as
 14  0 |  a = SI
  3  8 |  b = NO
```

Figura 9. Resultado BayesNet.

Se construyó el modelo a partir de 174 instancias las cuales se utilizaron como conjunto de entrenamiento y poder probar el modelo obtenido. Se aplicó un modelo de agrupamiento al conjunto de datos para la caracterización de los estudiantes admitidos, crear los distintos perfiles de los estudiantes en los diferentes grupos hallados y decidir qué otros factores determinan la separación de grupos producida por el algoritmo *k-Means*.

10. Conclusiones

Las metodologías ágiles para el desarrollo de software como SCRUM permiten tener la capacidad de contar con equipos humanos autogestionados, con alta motivación y un buen manejo de la creatividad. Además, esta metodología se ajusta a los cambios que se presentan durante las fases del proyecto. En la actualidad los proyectos de software se están encaminando hacia la generación de conocimiento mediante el uso de la información almacenada en las bases de datos institucionales; tomando como referente esta fuente de información para mejorar los procesos y solucionar los problemas que son objeto de estudio.

Se diseñó y construyó una herramienta software que permite recolectar información personal, académica, familiar y financiera de los estudiantes. Los datos obtenidos son estudiados mediante técnicas de minería de datos con el fin de generar reportes estadísticos y nuevas estrategias o programas de acompañamiento que apoyen los procesos de permanencia y graduación de los estudiantes de UNISANGIL.

La minería de datos es una técnica que sirve para hacer análisis y encontrar información útil para las organizaciones. En educación superior se aplica de varias maneras, siendo el problema de la deserción una de las áreas de estudio que con frecuencia es intervenido con la aplicación de diversos algoritmos y técnicas de agrupamiento de datos.

Al realizar las ejecuciones de los algoritmos seleccionados con los datos obtenidos en las entrevistas aplicadas a los desertores, se evidencia que el algoritmo de árbol de decisión demostró un mejor desempeño en comparación con el algoritmo de BayesNet. Esto se deduce a partir del mayor porcentaje de instancias correctamente clasificadas.

Para el caso de estudio (Programa de Enfermería, Sede San Gil), se trabajó con el algoritmo J48, el cual es una implementación del algoritmo C4.5, uno de los más utilizados en minería de datos en temas relacionados con la deserción. En el campo de la minería de datos, se puede aplicar modelos de aprendizaje automático con el fin de llegar a un modelo predictivo; también se debe continuar explorando las variables contenidas en la bodega de datos con Pentaho, como herramienta de *Business Intelligence,* aplicando el proceso de extracción, transformación y carga de los datos, para aprovechar las diferentes técnicas de minería que ofrece la herramienta y generar reportes gráficos, tableros de mando, para conocer el momento en el que se encuentran los estudiantes y, en consecuencia, definir nuevas estrategias de acompañamiento para los estudiantes.

Agradecimientos

Se agradece al Fondo para el Desarrollo de la Educación Superior – FODESEP por la financiación al proyecto de investigación FCNI-2017-S16 titulado "Diseñar e implementar un software que apoye el programa de permanencia y graduación estudiantil de UNISANGIL, a través de la generación de alertas tempranas y reportes para el análisis de datos", bajo la dirección de la ingeniera Luz Yamile Caicedo Chacón, docente investigador del Grupo de estudios avanzados en tecnologías de información y comunicación de UNISANGIL – HYDRA. También se agradece al Departamento de Investigación y al Departamento de Sistemas por el acompañamiento y disposición de recursos para poder contar con una aplicación que está alojada en los servidores institucionales.

Declaración de conflicto de intereses

Los autores declaran no tener conflicto de intereses con respecto a la investigación, autoría y/o publicación de este artículo.

Referencias

Ballesteros Román, A., Sánchez-Guzmán, D., y García Salcedo, R. (2013). Minería de datos educativa: Una herramienta para la investigación de patroness de aprendizaje sobre un contexto educativo. *Latin-American Journal of Physics Education,* 7(4), pp. 662–668. Recuperado de http://www.lajpe.org/dec13/22-LAJPE_814_bis_Alejandro_Ballesteros.pdf

Cárdenas Parra, C. N., Müller Rueda, J. S. y Ortiz Bernal, J. T. (2019). Detección de patrones de deserción estudiantil universitaria utilizando técnicas de análisis inteligente de datos. Fundación Universitaria de San Gil – UNISANGIL.

Figueroa, R. G., Solis, C. J. y Cabrera, A. A. (2008). Metodologías tradicionales vs. metodologías ágiles. Universidad

Técnica Particular de Loja, Escuela de Ciencias de la Computación.

Galán Cortina, V. (2015). Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario. Universidad Carlos III de Madrid. Recuperado de https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

Jiménez Galindo, Á. y Álvarez García, H. (2010). Minería de Datos en la Educación. Universidad Carlos III de Madrid.

Lara Gutiérrez, H. G., Lara Ruíz, M. G., Hernández Hernández, V., Hernández Hernández, B. y Hernández Hernández, G. (2016). Análisis de un caso práctico aplicando el algoritmo de k-Means mediante WEKA (Waikato Environment for Knowledge Analysis). Recuperado en septiembre 15 del 2017 de https://repository.uaeh.edu.mx/revistas/index.php/huejutla/article/download/1135/4718?inline=1

Mariño, S. I. y Alfonzo, P. L. (2014). Implementación de SCRUM en el diseño del proyecto de Trabajo Final de Aplicación. Scientia Et Technica, 19(4), pp. 413–418. Recuperado de: https://www.redalyc.org/pdf/849/84933912009.pdf

Ministerio de Educación Nacional. (2015). Guía para la implementación del modelo de gestión de permanencia y graduación estudiantil en Instituciones de Educación Superior. Bogotá D.C. Recuperado de https://www.mineducacion.gov.co/1759/articles-356272_recurso.pdf

Riveros Garzón, J. E. (2016). Deserción estudiantil Facultad de Ingeniería de Petróleos en la Fundación Universitaria de América. Universidad Militar Nueva Granada. Recuperado de https://repository.unimilitar.edu.co/bitstream/handle/10654/14758/RiverosGarzonJerryErnesto2016.pdf

Rodríguez Montequín, M. T., Álvarez cabal, J. V., Mesa Fernández, J. M. y González Valdés, A. (2005). Metodologías para la realización de proyectos de Data Mining. In *VII Congreso Internacional de Ingeniería de Proyectos* (pp. 257–265). Recuperado de https://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf

Rodríguez, O. (2016). Método k-medias. Recuperado de http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Presentación_-_k-means.13775252.pdf

Santiesteban Rojas, J. C. (2012). Definición de Redes Bayesianas y sus aplicaciones. Recuperado en septiembre 15 del 2017 de: http://vinculando.org/articulos/redes-bayesianas.html

Trigas Gallego, M. (2012). Metodología SCRUM. Desarrollo detallado de la fase de aprobación de un proyecto informático mediante el uso de metodologías ágiles. Recuperado de: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/17885/1/mtrigasTFC0612memoria.pdf

Vizcaíno Garzón, P. A. (2008). Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante el uso de Weka (Waikato Environment for Knowledge Analysis). Fundación Universitaria Konrad Lorenz. Recuperado de: http://www.konradlorenz.edu.co/images/stories/suma_digital_sistemas/2009_01/final_paula_andrea.pdf

Sobre los autores

Caicedo Chacón Luz Yamile

Ingeniero de Sistemas de la Universidad Autónoma de Bucaramanga, Especialista en Pedagogía de la Virtualidad de la Fundación Universitaria Católica del Norte, Máster en Business Intelligence Universitat de Barcelona / Escuela de Administración de Empresas – Online Business School (UB/EAE-OBS). Actualmente se desempeña como docente investigador de tiempo completo de la Facultad de Ciencias Naturales e Ingeniería de la Fundación Universitaria de San Gil – UNISANGIL. Entre sus áreas de interés se encuentran: diseño y desarrollo de bases de datos y bodegas de datos, desarrollo de software y aplicaciones con *business intelligence* y minería de datos.

Cárdenas Parra Cristian Noé

Ingeniero de sistemas de la Fundación universitaria de San Gil – UNISANGIL. Actualmente se desempeña como desarrollador de software. Entre sus áreas de interés se encuentran en desarrollo de software, manejo de base de datos, minería de datos y programación en dispositivos móviles.

Müller Rueda Juan Sebastián

Ingeniero de sistemas de la Fundación universitaria de San Gil – UNISANGIL. Actualmente se desempeña como desarrollador de software. Entre sus áreas de interés se encuentran la gestión y administración de base datos, así como la programación.

Ortiz Bernal Jeniffer Tatiana

Ingeniera de sistemas de la Fundación universitaria de San Gil – UNISANGIL. Actualmente se desempeña como desarrolladora de software. Entre sus áreas de interés, se encuentran el diseño y desarrollo de software, junto con el análisis de datos y el manejo de base de datos.