UTN.BA

UTN.BA
ESCUELA DE
POSGRADO

Trabajo Final de Especialización en

Ingeniería en Sistemas de Información

"Elementos para una Propuesta de Métricas para Proyectos de Explotación de Información para PyMEs"

Tesista: Ing. Diego M. Basso

Director

Mg. Darío Rodríguez
Profesor Adjunto Regular
Laboratorio de Investigación y Desarrollo en Ingeniería de Explotación de Información
Universidad Nacional de Lanús

Co-Director

Mg. Eduardo Diez
Profesor Asociado
Laboratorio de Investigación y Desarrollo en Aseguramiento de la Calidad de Software
Universidad Nacional de Lanús

Ciudad Autónoma de Buenos Aires, 2014

RESUMEN

Los Proyectos de Explotación de Información requieren de un proceso de planificación para estimar el esfuerzo, el tiempo y medir diferentes aspectos del producto para garantizar la calidad del mismo. Los procesos de desarrollo tradicionales y las métricas usuales de la Ingeniería de Software y la Ingeniería del Conocimiento no son totalmente adecuados para estos proyectos, ya que las etapas de desarrollo y los parámetros utilizados son de naturaleza y características diferentes. Por consiguiente, existe la necesidad de tener métricas específicas para los Proyectos de Explotación de Información, con particular énfasis en las características de las empresas PyMEs. En ese contexto, se describe un marco conceptual para la definición de una propuesta de métricas aplicables al desarrollo de este tipo de proyectos, siguiendo los lineamientos del Modelo de Proceso de Desarrollo para Proyectos Explotación de Información.

ABSTRACT

Information Mining Projects require a planning process to estimate the effort, time and measure different aspects of the product to ensure its quality. Traditional development processes and the usual metrics of Software Engineering and Knowledge Engineering are not considered adequate to these projects, as the development stages and the parameters used are of different nature and characteristics. There is therefore the need to have specific metrics for Information Mining Projects, with particular emphasis on the characteristics of SMEs (SMEs=PyMEs for its acronym in Spanish) companies. In this context, a conceptual framework is described for the definition of a proposal metrics applicable to the development of such projects, along the lines of the Model Development Process for Information Mining Projects.

ÍNDICE

1. INTRODUCCIÓN 1

1.1. Objetivos 1

1.1.1. Objetivo General 1

1.1.2. Objetivos Específicos 1

1.2. Fundamentos del Trabajo 2

1.3. Metodología Empleada 3

1.4. Visión General del Trabajo 3

1.5. Producción Científica Vinculada a este Trabajo 4

2. ESTADO DE LA CUESTIÓN 5

2.1. Importancia de las Métricas en el Desarrollo de Proyectos 5

2.1.1. Definición de Métrica 6

2.1.2. Atributos Medibles en el Software 7

2.2. Métricas Existentes 8

2.2.1. Métricas en Ingeniería de Software 9

2.2.2. Métricas en Ingeniería del Conocimiento 11

2.3. Contexto de los Proyectos de Explotación de Información 12

2.3.1. Estimación del Esfuerzo en Proyectos de Explotación de Información 14

2.3.2. Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información 16

2.3.3. Procesos de Explotación de Información 18

2.4. Necesidad de Métricas en Proyectos de Explotación de Información 20

2.4.1. Métricas Existentes aplicables a Proyectos de Explotación de Información 20

3. CONCLUSIONES 23

3.1. Resumen de los Resultados del Trabajo 23

3.2. Futuras Líneas de Investigación 23

4. REFERENCIAS 25

ÍNDICE DE FIGURAS

Figura 2.1 Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información 16

ÍNDICE DE TABLAS

Tabla 2.1 Métricas existentes en Ingeniería de Software 10
Tabla 2.2 Características del Método de Estimación de Esfuerzo DMCoMo 15
Tabla 2.3 Características del Método de Estimación de Esfuerzo para PyMEs 16
Tabla 2.4 Tareas vinculadas al Modelo de Proceso de Desarrollo 18

NOMENCLATURA Y ACRÓNIMOS

CMM	Capability Maturity Model. Modelo de Madurez de Capacidades
CMMI	Capability Maturity Model Integration. Integración de Modelos de Madurez de Capacidades
COCOMO	Constructive Cost Model. Modelo Constructivo de Costos
CRISP-DM	Metodología para el Desarrollo de Proyectos de Explotación de Información
DM	Data Mining. Minería de Datos
DMCoMo	Data Mining Cost Model. Modelo de Estimación para Proyectos de Explotación de Información
DW	Data Warehouse. Almacén de Datos.
IEEE	Institute of Electrical and Electronics Engineers. Instituto de Ingeniería Eléctrica y Electrónica
IFPUG	International Function Point Users Group. Grupo Internacional de Usuarios de Puntos Función
IM	Information Mining. Explotación de Información
ISO	International Organization for Standardization. Organización Internacional de Normalización
P^3TQ	Metodología para el Desarrollo de Proyectos de Explotación de Información (Product, Place, Price, Time, Quantity)
PyMEs	Pequeñas y Medianas Empresas
SEI	Software Engineering Institute. Instituto de Ingenería de Software
SEMMA	Metodología para el Desarrollo de Proyectos de Explotación de Información (Sample, Explore, Modify, Model and Assess)
SQA	Aseguramiento de la Calidad del Software

"Cuando puedes medir aquello de lo que estás hablando y expresarlo en números, sabes algo sobre ello; pero cuando no lo puedes medir y no lo puedes expresar en números, tu conocimiento acerca de ello es insatisfactorio".

William Thomson (Lord Kelvin)

1. INTRODUCCION

En este capítulo se presentan los objetivos de este trabajo (sección 1.1), en particular el objetivo general (sección 1.1.1) y los objetivos específicos (sección 1.1.2), se presentan los fundamentos que conllevan este trabajo (sección 1.2) y se describe la metodología empleada para su desarrollo (sección 1.3). El capítulo finaliza con la descripción de la visión general del trabajo (sección 1.4) y el resumen de las producciones científicas vinculadas al mismo (sección 1.5).

1.1. Objetivos

Se dividen los objetivos de esta tesis en un objetivo general a alcanzar (sección 1.2.1), un conjunto de objetivos específicos (sección 1.2.2) que definen los pasos a seguir para lograr el objetivo general y el alcance previsto con estos objetivos (sección 1.2.3).

1.1.1. Objetivo General

El objetivo de este trabajo es construir el estado del arte para la definición de una propuesta de métricas aplicables al proceso de desarrollo de Proyectos de Explotación de Información basado en el Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información para PyMEs (Pequeñas y Medianas Empresas).

1.1.2. Objetivos Específicos

Se detallan a continuación los objetivos específicos que permiten, en conjunto, establecer los pasos a seguir para lograr cumplir con el objetivo general:

- Diferenciar las métricas existentes aplicables a la construcción tradicional de software y las utilizadas para medir la madurez del conocimiento y complejidad del dominio en el desarrollo de sistemas expertos.

- Determinar los límites y alcances del Modelo de Proceso de Desarrollo para Proyectos Explotación de Información para PyMEs, identificando los subprocesos y tareas del modelo que servirían de guía para la definición de una propuesta de métricas aplicables a estos proyectos.
- Establecer las diferencias entre el modelo de estimación de esfuerzo para proyectos de Data Mining DMCoMO y el método de estimación para proyectos de Explotación de Información propuesto para PyMEs, identificando los factores de costo considerados por ambos métodos, y que servirían de referencia para la clasificación de las propuesta de métricas para Proyectos de Explotación de Información.
- Identificar métricas existentes del ámbito de la Ingeniería de Software y la Ingeniería del Conocimiento que serían de aplicación a los Proyectos de Explotación de Información.

1.2. Fundamentos del Trabajo

El proceso de planificación de todo proyecto de software debe hacerse partiendo de una estimación del trabajo a realizar. Para obtener software de calidad es preciso medir el proceso de desarrollo, cuantificar lo que se ha hecho y lo que falta por hacer, estimar el tamaño del proyecto, costos, tiempo de desarrollo, control de calidad, mejora continua y otros parámetros. En este sentido, las métricas ayudan a entender tanto el proceso que se utiliza para desarrollar un producto, como el propio producto. Asimismo, tienen un papel decisivo en la obtención de un producto de alta calidad, porque determinan mediante estadísticas basadas en la experiencia, el avance del software y el cumplimiento de parámetros requeridos.

En la Ingeniería de Software, los proyectos de desarrollo tradicionales aplican una amplia diversidad de métricas e indicadores para especificar, predecir, evaluar y analizar distintos atributos y características de los productos y procesos que participan en el desarrollo y mantenimiento del software. La aplicación de un enfoque cuantificable es una tarea compleja que requiere disciplina, estudio y conocimiento de las métricas e indicadores adecuados para los distintos objetivos de medición y evaluación, con el fin de garantizar la calidad del software construido.

En el ámbito de la Ingeniería en Conocimiento, especialmente en el desarrollo de sistemas expertos o sistemas basados en conocimientos, un aspecto importante que debe ser medido es la conceptualización para poder estimar actividades futuras y obtener información del estado de madurez del conocimiento sobre el dominio y sus particularidades [Hauge *et al.*, 2006]. Estas métricas de madurez de conceptualización para Sistemas Expertos definidas en [Hauge *et al.*, 2006]

y aplicadas en [Pollo-Cattaneo, 2007; Pollo-Cattaneo *et al.*, 2008] brindan además información sobre la complejidad del dominio.

Los Proyectos de Explotación de Información también requieren de un proceso de planificación que permita estimar sus tiempos y medir el avance del producto en cada etapa de su desarrollo y calidad del mismo. Sin embargo, como consecuencia de las diferencias que existen entre un proyecto clásico de construcción de software y un proyecto de explotación de información [Vanrell *et al.*, 2010a], las métricas de software usuales no se consideran adecuadas ya que los parámetros utilizados en los proyectos de explotación de información son de naturaleza diferentes [Marbán, 2003; Marbán *et al.*, 2008] y no se ajustan a sus características específicas.

1.3. Metodología Empleada

Mediante este trabajo se pretende establecer un marco teórico para realizar una propuesta de métricas que puedan ser utilizables en el desarrollo de Proyectos de Explotación de Información. En tal sentido, se describen los siguientes pasos metodológicos:

La primera etapa consiste en describir la importancia de la utilización de métricas en la gestión de proyectos de desarrollo de software, establecer una clasificación de las métricas en relación a los aspectos del software que miden, e identificar métricas existentes de la Ingeniería de Software y la Ingeniería del Conocimiento de aplicación para proyectos de Explotación de Información.

En una segunda etapa y en base a la investigación documental, se mencionan las características consideradas por los métodos de estimación de esfuerzo definidos para proyectos de explotación de información, y se identifican los límites, subprocesos y tareas asociadas al Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información. Asimismo se describen los distintos procesos de explotación de información basados en sistemas inteligentes.

Finalmente se elabora el informe final con las conclusiones obtenidas el cual se tomará como base para la realizar una Propuesta de Métricas para Proyectos de Explotación de Información, con especial interés en proyectos de empresas PyMEs.

1.4. Visión General del Trabajo

Este trabajo se encuentra dirigido principalmente a profesionales, estudiantes y cátedras universitarias vinculadas a la Ingeniería de Software y la Inteligencia de Negocios, que se dediquen al estudio y desarrollo de métodos y herramientas para el campo de la disciplina en Ingeniería de Proyectos de Explotación de Información. El mismo se estructura en cinco capítulos: Introducción,

Estado de la Cuestión, Necesidad de Métricas para Proyectos de Explotación de Información, Conclusiones y Referencias, los cuales se describen a continuación.

En el Capítulo 1 – "Introducción" se plantea el contexto que da soporte a este trabajo, se establecen los objetivos, en particular el objetivo general y los objetivos específicos y se describe la metodología empleada para el desarrollo del trabajo. El capítulo finaliza con la descripción de la visión general del trabajo.

En el Capítulo 2 – "Estado de la Cuestión" se plantea la importancia que tienen las métricas en el desarrollo de proyectos y para el aseguramiento de la calidad, se clasifican las métricas en base a los aspectos que pueden medirse en el software y se identifican diversas métricas utilizadas en la Ingeniería de Software y la Ingeniería del Conocimiento, especialmente en el desarrollo de los Sistemas Expertos. Posteriormente, se describe el contexto de la Ingeniería de Proyectos de Explotación de Información, identificando las diferencias de esta disciplina con la Ingeniería del Software. Asimismo, se mencionan las particularidades de los métodos propuestos para estimación de esfuerzo en proyectos de explotación de información y los procesos definidos para realizar esta explotación y se describe un Modelo de Procesos de Desarrollo para estos proyectos. Luego, se enuncian algunas conclusiones parciales consideradas dentro del marco de una propuesta de métricas específicas, y se identifican aquellas métricas de la Ingeniería de Software y la Ingeniería del Conocimiento que podrían considerarse de aplicación a los proyectos de Explotación de Información.

En el Capítulo 3 – "Conclusiones" se mencionan las conclusiones obtenidas a partir del desarrollo del marco conceptual, estableciendo la línea base para realizar una propuesta de métricas específicas para los proyectos de interés de este trabajo.

En el Capítulo 4 – "Referencias" se enuncian las referencias bibliográficas utilizadas para este trabajo.

1.5. Producción Científica Vinculada a este Trabajo

Durante el desarrollo de este trabajo se realizó la comunicación de resultados parciales a través de una publicación al congreso que se menciona a continuación:

- Basso, D., Rodríguez, D., García-Martínez, R. (2013). *Propuesta de Métricas para Proyectos de Explotación de Información*. Workshop de Bases de Datos y Minería de Datos. Proceedings XIX Congreso Argentino de Ciencias de la Computación. Pag. 983-992. ISBN 978-987-23963-1-2.

2. ESTADO DE LA CUESTIÓN

Este capítulo describe cómo es el contexto de trabajo en cual se inserte este trabajo, permitiendo al lector adquirir los conocimientos necesarios para comprender la problemática de la misma. En el mismo se indica la importancia que tienen las métricas en el desarrollo de proyectos (2.1) y su relación con la calidad, se identifican algunas métricas existentes (sección 2.2) en el ámbito de la Ingeniería de Software y la Ingeniería del Conocimiento, en especial en el desarrollo de los Sistemas Expertos y se continúa con una contextualización de la Ingeniería de Proyectos de Explotación de Información (sección 2.3). El capítulo finaliza planteando la necesidad de reunir métricas para estos proyectos (sección 2.4) y presentando algunas conclusiones parciales obtenidas de investigaciones anteriores sobre el tema.

2.1. Importancia de las Métricas en el Desarrollo de Proyectos

En la actualidad, existe un importante interés por parte de las empresas que desarrollan software por lograr que los productos software cumplan con ciertos indicadores de calidad en todas las etapas del desarrollo [Porta García *et al.*, 2012]. Con independencia del tipo de producto que se desarrolle en un proyecto, la calidad es fundamental para lograr la satisfacción de las necesidades y expectativas del cliente. En [Pressman, 2005] se menciona que el aseguramiento de la calidad del software (SQA) es una "actividad de protección" que se aplica a lo largo de todo el proceso de Ingeniería de Software, en la que se incluyen mecanismos para medir el producto y el proyecto, entre otros. #

Dentro de los estándares que proporcionan modelos para la evaluación de la calidad del software, se encuentran las normas ISO 9000 (especialmente ISO 9001 e ISO 9003-2) [Rodríguez *et al.*, 1998], CMM (Capability Maturity Model) [Sanders y Curran, 1995] y CMMI (Capability Maturity Model Integration) [SEI, 2010]. Estos modelos incorporan técnicas y procesos para el aseguramiento de la calidad que se corresponden con la medición del software. Por otra parte, algunos autores [Pressman, 2005; McCall *et al.*, 1977] y estándares internacionales [ISO/IEC 9126-1, 2001; ISO/IEC 25010, 2011], han tratado de determinar y categorizar las características que deben cumplir todo producto software para ser considerado de calidad, y a partir de éstas proporcionar la terminología para especificar, medir y evaluar la calidad del mismo.

La medición de un software es tan importante en cualquier proceso de ingeniería como su misma construcción [Estayno *et al.*, 2009], ya que permite tener una visión del proyecto, de la evaluación del producto y de su nivel de aceptación, logrando un mejoramiento continuo del software y

permitiendo cuantificar y gestionar, de forma más efectiva, cada una de las variables a las que se necesite hacer seguimiento. En este sentido, la medición del software debe satisfacer tres objetivos fundamentales [Fenton y Pfleeger, 1997]: (1) ayudar a entender qué ocurre durante el desarrollo y el mantenimiento, (2) permitir controlar qué es lo que ocurre en los proyectos y (3) poder mejorar los procesos y productos. Mejorar la calidad de los resultados de un proyecto de software o la eficiencia de sus procesos es difícil, si no se recolectan métricas.

En el ámbito de la Ingeniería de Software y la Ingeniería del Conocimiento existen métricas e indicadores, que comprenden un conjunto de actividades en el desarrollo de un proyecto de software (entre los que se incluye el aseguramiento y control de calidad), las cuales permiten analizar y evaluar características y atributos de los productos y procesos que participan en el desarrollo.

Los proyectos de Explotación de Información también deben considerar la aplicación de una metodología de desarrollo [García-Martínez *et al.*, 2011] que incluya entre sus actividades el registro de métricas, que permitan medir y controlar el avance del proyecto y evaluar su calidad. Asimismo, se destaca la necesidad de realizar estimaciones de esfuerzo al comienzo de estos proyectos y compararlos con los valores reales al finalizar el mismo.

Dentro del cuerpo de conocimiento de la Ingeniería de Proyectos de Explotación de Información, se han propuesto y desarrollado distintas herramientas [García-Martínez *et al.*, 2011], entre las que se mencionan: un modelo de procesos, un proceso de educción de requisitos, un método de estimación, una metodología de selección de herramientas, un proceso de transformación de datos y una serie de procesos basados en técnicas de minería de datos. Estas herramientas además, han sido utilizadas en proyectos para pequeños y medianos emprendimientos. Asimismo, en el trabajo de [García-Martínez *et al.*, 2011] se ha señalado la necesidad de plantear métricas significativas asociadas al proceso de desarrollo de un proyecto de Explotación de Información, que permitan suministrar información relevante a tiempo y establecer objetivos de mejora en los procesos y productos, con el fin de garantizar la calidad de estos proyectos.

Se presenta continuación la definición del concepto de métrica (sección 2.1.1) y su relación con las medidas, indicadores y medición, y se identifican los atributos que se pueden medir en un proyecto de desarrollo de software (sección 2.1.2).

2.1.1. Definición de Métrica

Una métrica se puede definir básicamente como la medición numérica de un atributo ante la necesidad de tener información cuantitativa del mismo para la toma de decisiones.

En los proyectos de desarrollo de software, a menudo se suele hablar de *métricas* y de *medidas*, indistintamente. Sin embargo, existen diferencias entre estos términos. Se proporcionan a continuación algunas definiciones que permiten entender el concepto de métrica.

- **Medida**: valor asignado a un atributo de una entidad mediante una medición. Es una medida que proporciona una indicación cuantitativa de extensión, cantidad, dimensiones, capacidad y tamaño de algunos atributos de un proceso o producto [Pressman, 2005].
- **Métrica**: medida cuantitativa del grado en que un sistema, componente o proceso posee un determinado atributo [IEEE, 1993; Pressman, 2005].
- **Indicador**: es una métrica o una combinación de métricas que proporcionan una visión profunda del proceso del software, del proyecto de software, o del producto en sí [Pressman, 2005].
- **Medición**: es el proceso por el cual los valores son asignados a atributos o entidades en el mundo real tal como son descritos de acuerdo a reglas claramente definidas [Fenton y Neil, 1999]. Dicho de otro modo, es el proceso por el cual se obtiene una medida [Pollo-Cattaneo, 2007].

Estas definiciones permiten afirmar que a partir de los valores de las medidas, es posible reunir métricas que proporcionen información mediante indicadores, para poder controlar la eficacia del proceso, del proyecto o del producto software [Pollo-Cattaneo, 2007].

2.1.2. Atributos Medibles en el Software

Cualquier cosa que se quiera medir o predecir en un software representa un atributo (propiedad) de cualquier entidad de un producto, proceso o recurso asociado a éste. Cada entidad de software tiene varios atributos internos y externos que pueden ser medidos [Fenton y Pfleeger, 1997].
Los atributos internos de un producto, proceso o recurso son aquellos que se pueden medir directamente en términos del producto, proceso o recurso del mismo [Fenton y Neil, 1999], por ejemplo: el tamaño del software, el esfuerzo para desarrollar un módulo del software, el tiempo transcurrido en la ejecución de cualquier módulo de software, entre otros. Los atributos externos de un producto, proceso o recurso son aquellos que solamente pueden ser medidos con respecto a cómo el producto, proceso o recurso se relacionan con su entorno [Pressman, 2005], por ejemplo: el costo de eficacia de los procesos, productividad del grupo de desarrollo, complejidad del proyecto, la usabilidad, fiabilidad, o portabilidad de un sistema, entre otros. Los atributos externos son los más difíciles de medir, porque estos no pueden ser medidos directamente [Fenton y Neil, 1999].

Los valores de los atributos se obtienen tras realizar mediciones sobre el software. Las mediciones dan como resultado una serie de métricas, que según la norma ISO/IEC 9126 [ISO/IEC 9126-2, 2003; ISO/IEC 9126-3, 2003; ISO/IEC 9126-4, 2004] se pueden clasificar en tres categorías, según sea su naturaleza:

- **Métricas básicas**: son métricas que se obtienen directamente del análisis del código o la ejecución del software. No involucra ningún otro atributo ni depende de otras métricas. En [Pressman, 2005] estas métricas se denominan *directas*. Entre las métricas básicas se tiene la cantidad de líneas de código del programa o de cada módulo, la cantidad de horas de desarrollo, la cantidad de fuentes de datos o tablas a utilizar, la cantidad de atributos y registros de una tabla, entre otras.
- **Métricas de agregación**: son métricas compuestas a partir de un conjunto definido de métricas básicas (o directas), generalmente mediante una suma ponderada.
- **Métricas derivadas**: son métricas compuestas por una función de cálculo matemático, que utiliza como variables de entrada el valor de otras métricas. En [Pressman, 2005] estas métricas se denominan *indirectas*. Entre las métricas derivadas se tiene la cantidad de líneas de código producidas por hora y por persona, el porcentaje de completitud del proyecto, el tamaño promedio de los módulos del software, el tiempo promedio que una persona dedica a corregir los defectos de un módulo, entre otras.

2.2. Métricas Existentes

Para medir el desarrollo de un proyecto es necesario saber qué entidades son medidas y tener una idea de los atributos de la entidad. Para ello, se debe identificar el atributo a medir y su significado de medición [Negro, 2008]. Por otra parte, diversos autores han propuesto distintos tipos de métricas de acuerdo a la relevancia de lo que se esté midiendo.

Para poder llegar a un buen resultado en un proyecto de software se debe comprender el dominio del problema donde se va a trabajar, los recursos necesarios, las actividades y tareas a llevar a cabo, el esfuerzo y tiempo que se va a insumir, el plan de acción y los riesgos que se van a correr [Pollo-Cattaneo, 2007]. En este sentido, la naturaleza, el tamaño del proyecto y el entorno en el que se desarrolla son factores determinantes y afectan en gran medida a la estimación que se realice.

Dentro del campo de la Ingeniería de Software (sección 2.2.1) e Ingeniería del Conocimiento (sección 2.2.2) existen diferentes tipos de métricas, que pueden ser clasificadas según los aspectos que miden.

2.2.1. Métricas en Ingeniería de Software

Las métricas de software proporcionan información relevante a tiempo que contribuye a gestionar de forma más efectiva un proyecto, y mejorar la calidad de los procesos y productos de software [Pressman, 2005]. Además, al conocerse el estado actual del desarrollo de un proyecto, pueden establecerse objetivos de mejora [Kan *et al.*, 2001].

Por otra parte, el uso de métricas no sólo permite entender, monitorizar, controlar, predecir y probar el desarrollo de software y los proyectos de mantenimiento [Briand *et al.*, 1998] sino que también pueden ser utilizadas para tomar mejores decisiones [Pfleeger, 1997].

En el campo de la Ingeniería de Software, es sabido que contar con datos históricos de proyectos terminados, contribuye a estimar con mayor exactitud el esfuerzo, tiempo de desarrollo, costo, posibles errores, recursos y tamaño para los nuevos proyectos, facilitando las tareas de planificación, seguimiento y control del mismo. Esto implica que las métricas se consideran necesarias y de gran importancia, ya que proporcionan información objetiva que contribuye al mejoramiento de los procesos y productos de software, favoreciendo al logro de la calidad y a una posterior evaluación del nivel de satisfacción del usuario. En este contexto, los proyectos de desarrollo tradicionales aplican diversas métricas cuantitativas e indicadores, en todas y cada una de las fases del ciclo de vida del software (especificación, análisis, diseño, construcción, pruebas y documentación).

Las métricas de software contemplan varias clasificaciones que apuntan a diferentes aspectos del proceso y del producto de software [Pressman, 2005]. De acuerdo al contexto o dominio de aplicación y de las características o atributos del software, las métricas de software se pueden tipificar en: métricas del producto, del proceso y del proyecto. A su vez, algunas de estas métricas pueden pertenecer a más de una clasificación.

- **Métricas del producto**: son métricas que evalúan la calidad de los productos entregables, permitiendo tener un conocimiento detallado del diseño y la construcción del producto software. En estas métricas se tienen en cuenta atributos como: tamaño, calidad, complejidad, esfuerzo, volatilidad, entre otros.
- **Métricas del proceso**: son métricas aplicadas a fines estratégicos y propician indicadores que conducen a avances en el proceso y ambiente de desarrollo del software, a partir de información histórica de procesos similares. Se utilizan para evaluar si la eficiencia de un proceso ha mejorado en el largo plazo. Se recopilan de todos los proyectos y durante un largo período de tiempo. Dentro de estas métricas se incluye atributos como la experiencia del grupo, el costo del

desarrollo y mantenimiento, el esfuerzo y tiempo dedicado a las pruebas, tiempo de desarrollo (total y por proceso, subproceso), tipo y cantidad de fallas, número de cambio con modelos previos, costo de aseguramiento de la calidad, cantidad de personas por día, por mes, intensidad del trabajo, interrupciones, entre otros [Screpnik, 2013].

- **Métricas del proyecto**: son métricas de tipo tácticas y describen las características propias del proyecto y de su ejecución. Estas métricas reducen la planificación del desarrollo ya que permite realizar los ajustes necesarios para evitar retrasos o riesgos potenciales, minimizar los defectos y por lo tanto la cantidad de trabajo que debe rehacerse, ocasionando en consecuencia una reducción del costo global del proyecto [McDermid, 1991]. A su vez, permiten evaluar la calidad de los productos obtenidos en cada etapa del desarrollo [McDermid, 1991]. Estas métricas tienen en cuenta atributos como duración real del proyecto, esfuerzo real [persona-mes] por proceso, subproceso y por proyecto, progreso del proyecto, tamaño del proyecto, costo total invertido, entre otros.

Como se mencionó, diversos autores (Boehm, Albretch, McCall, Pressman, entre otros), estándares y normas internacionales (IFPUG, IEC/ISO 9126, IEEE, entre otras) han propuesto un amplio conjunto de métricas de software aplicables al campo de la Ingeniería de Software. En la tabla 2.1 se menciona un grupo acotado de métricas existentes en esta disciplina, de acuerdo al aspecto y atributo del software que miden.

Aspectos del Software	Atributo	Métricas Existentes
Producto	Tamaño	▪ Líneas de Código (medidas en miles - KLDC) ▪ Puntos de Función (PF) ▪ Páginas de Documentación
	Complejidad	▪ Complejidad ciclomática ▪ Nivel de acoplamiento de los módulos ▪ Nivel de modularidad (cohesión de módulos)
	Calidad	▪ Cantidad de defectos por KLDC ▪ Cantidad de errores encontrados por KLDC ▪ Cantidad de defectos/errores que encuentran los usuarios después de la entrega ▪ Tipo y origen de los defectos (requerimientos, análisis y diseño, construcción, integración y pruebas)
	Mantenimiento	▪ Cantidad de componentes ▪ Volatilidad de los componentes ▪ Complejidad de los componentes ▪ Cantidad de requerimientos nuevos, de cambios o mejoras ▪ Cantidad de requerimientos de corrección de defectos ▪ Tiempo promedio de corrección de errores ó defectos ▪ Tiempo promedio de cambios ▪ Porcentaje del código corregido
	Confiabilidad	▪ Tiempo transcurrido entre fallas ▪ Tiempo esperado entre fallas ▪ Tiempo requerido para corregir una falla ▪ Nivel de severidad de la falla
	Usabilidad	▪ Facilidad de aprendizaje de uso

Aspectos del Software	Atributo	Métricas Existentes
		▪ Errores cometidos por los usuarios con el uso ▪ Tiempo requerido para realizar las tareas
	Rendimiento	▪ Tiempos de respuesta (acuerdos SLA) ▪ Utilización de recursos (Troughput o Thruput) – cantidad de transacciones que pueden ejecutarse concurrentemente con un tiempo de respuesta razonable. ▪ Tiempo de recuperación
Proyecto	Esfuerzo	▪ Cantidad de horas trabajadas ▪ Cantidad de personas que trabajan en el proyecto ▪ Tiempo transcurrido ▪ Distribución del esfuerzo por fase
	Costo	▪ Costo del Desarrollo ▪ Costo del Soporte ▪ Costo de hs/persona
	Productividad	▪ Cantidad de software desarrollado por unidad de tiempo de trabajo ▪ Tamaño/Esfuerzo ▪ Ritmo de entrega del software por unidad de tiempo transcurrido.
	Seguimiento	▪ Cronograma real vs Cronograma estimado ▪ Porcentaje de tareas completadas ▪ Porcentaje de requerimientos implementados por unidad de tiempo ▪ Porcentaje de tiempo total dedicado a las pruebas ▪ Porcentaje de error en la estimación del tiempo ▪ Costo sobre el valor agregado
	Estabilidad	▪ Origen de los cambios en los requerimientos ▪ Cambios de los requerimientos en el desarrollo ▪ Cambios en los requerimientos en producción
Proceso	Esfuerzo	▪ Distribución del esfuerzo por fase del proceso ▪ Cantidad de personas requeridas ▪ Esfuerzo requerido para corregir un defecto ▪ Esfuerzo requerido para mejorar un defecto
	Reusabilidad	▪ Cantidad de componentes reutilizados ▪ Grado de reusabilidad de los componentes
	Calidad	▪ Cantidad de defectos sin corregir ▪ Costo de corrección de defectos ▪ Eficacia en la eliminación de defectos ▪ Cantidad de veces que un módulo fue probado ▪ Tamaño del módulo ▪ Tiempo promedio de corrección de defectos
	Soporte a los clientes	▪ Tamaño del back log de defectos ▪ Tiempo de respuesta en atender los defectos ▪ Tiempo de resolución de defectos
	Herramientas	▪ Soporte de herramientas para procesos propuestos

Tabla 2.1. *Métricas existentes en Ingeniería de Software [elaboración propia]*

2.2.2. Métricas en Ingeniería del Conocimiento

La planificación juega un papel esencial en la gestión de un proyecto de software, en la que se debe estimar el esfuerzo humano, costo y tiempo. Para esto se tienen métricas de software, que permiten obtener información y así generar conocimiento de la evolución y alcance del proyecto. En el desarrollo de Sistemas Expertos o Sistemas Basados en Conocimiento, la planificación presenta particularidades que la hacen altamente compleja [Pollo-Cattaneo *et al.*, 2008]. Como se mencionó, proceso y producto son elementos protagonistas de las técnicas de medición. En este sentido, en el trabajo de [Hauge *et al.*, 2006] se han propuesto un conjunto de métricas de madurez aplicables en

la fase de conceptualización, que examinan el dominio del problema en el contexto de desarrollo de un Sistema Experto [Firestone, 2004]. Estas métricas, aplicadas en [Pollo-Cattaneo, 2007; Pollo-Cattaneo *et al.*, 2008], brindan además información sobre la madurez de la base de conocimientos y la complejidad del dominio. Se presentan a continuación las métricas existentes definidas en [Hauge *et al.*, 2006], las cuales se basan en *Reglas*, *Conceptos*, *Atributos* y *Niveles de Descomposición*.

- Número de Conceptos, Número de Reglas o Número de Atributos
- Número de Conceptos en una Regla / Número de Conceptos
- Número de Atributos en una Regla / Número de Atributos
- Número de Conceptos / Número de Reglas
- Número promedio de Atributos por Concepto
- A* (Número de Conceptos) + B* (Número de promedio de Atributos por Concepto)
- Número promedio de Niveles en un árbol de decisión
- Número promedio de Conceptos incluidos en cada Regla
- Número promedio de Atributos incluidos en cada Regla
- A* (Número promedio de Atributos en la Regla) + B* (Número de Reglas) + C* (Número promedio de Niveles de Descomposición)
- Número promedio de Reglas en las que se encuentra incluido cada Concepto
- A* Número promedio de Reglas por Conceptos que se encuentra incluido en B^+ Número de Conceptos
- Número promedio de Reglas en las que se encuentra incluido cada Atributo
- Para todos los niveles (Número de Decisiones en el Nivel i*i) / Número Total de Decisiones

2.3. Contexto de los Proyectos de Explotación de Información

La Explotación de Información (en inglés Information Mining, IM) es la sub-disciplina de los sistemas de información vinculada a la Inteligencia de Negocio [Negash y Gray, 2008] que aporta las herramientas de análisis y síntesis para extraer conocimiento, que se encuentra de manera implícita en los datos disponibles de diferentes fuentes de información [Schiefer *et al.*, 2004]. Dicho de otra manera, las herramientas que permiten transformar la información en conocimiento [Mobasher *et al.*, 1999; Srivastava *et al.*, 2000; Abraham, 2003; Cooley, 2003; García-Martínez *et al.*, 2011].

En [Larose, 2005] se define a la Explotación de Información como el proceso de descubrir nuevas correlaciones, patrones y tendencias significativas utilizando grandes cantidades de datos almacenados en repositorios, usando tecnologías de reconocimiento de patrones, así como técnicas matemáticas y de estadística. En este contexto, la Ingeniería de Proyectos de Explotación de

Información estudia los procesos de extracción de conocimiento no trivial [Martins, 2013], el cual es previamente desconocido y puede resultar útil para algún proceso [Stefanovic *et al.*, 2006].

Un proceso de Explotación de Información se define como un conjunto de tareas relacionadas lógicamente [Curtis *et al.*, 1992; García-Martínez *et al.*, 2013], el cual engloba un conjunto de técnicas de minería de datos (en inglés Data Mining, DM) que pueden ser elegidas para realizarlas y así lograr extraer de conocimiento procesable, implícito en el almacén de datos (en inglés Data Warehouse, DW) de la organización. Las bases de estas técnicas se encuentran en el análisis estadístico y en los sistemas inteligentes. De esta manera, se aborda la solución a problemas de predicción, clasificación y segmentación [Umapathy, 2007].

Un proyecto de Explotación de Información involucra, en general las siguientes fases [Maimon y Rokach, 2005]: comprensión del negocio y del problema que se quiere resolver, determinación, obtención y limpieza de los datos necesarios, creación de modelos matemáticos, ejecución, validación de los algoritmos, comunicación de los resultados obtenidos, e integración de los mismos, si procede, con los resultados en un sistema transaccional o similar. La relación entre todas estas fases tiene una complejidad que se traduce en una jerarquía de subfases.

Por otra parte, y a partir de la experiencia adquirida en proyectos de Explotación de Información, se han desarrollado diferentes metodologías de desarrollo que permiten gestionar esta complejidad de una manera uniforme, siendo CRISP-DM [Chapman *et al.*, 2000], SEMMA [SAS, 2008] y P^3TQ [Pyle, 2003] las metodologías probadas por la comunidad científica [Britos, 2008; Gambin y Pallota, 2009].

Los proyectos de desarrollo de software tradicional y de sistemas expertos necesitan un proceso de planificación que contemple un método de estimación de esfuerzo y tiempos, y la posibilidad de realizar el seguimiento y control del proyecto en cada fase de su desarrollo. Este seguimiento debe proveer de métricas e indicadores que permitan evaluar la calidad del proceso aplicado, el producto construido y los resultados obtenidos al finalizar el proyecto. En este sentido, los proyectos de Explotación de Información no escapan a esta misma necesidad.

Sin embargo, las fases habituales de un proyecto clásico de desarrollo de software (análisis, diseño, construcción, integración y prueba) no encuadran con las etapas propias de un proyecto de Explotación de Información [Vanrell *et al.*, 2010a]. Esto significa que las herramientas clásicas de la Ingeniería de Software tales como la ingeniería de requerimientos, los modelos de procesos, los ciclos de vida y los mapas de actividades no sean aplicables para los proyectos de Explotación de Información [García-Martínez *et al.*, 2011].

Por otra parte, un proyecto de Explotación de Información considera entre sus características más representativas la cantidad de fuentes de información a utilizar, el nivel de integración y calidad que presentan los datos, el tipo de problema de explotación de información a ser resuelto, los modelos necesarios a construir para el proyecto y la utilidad e interés de los patrones de conocimiento descubierto, entre otras. Estas características, muestran que las métricas de software y de sistemas expertos existentes, tampoco pueden considerarse del todo adecuadas, ya que los parámetros que utilizan estos proyectos son de naturaleza diferentes [Marbán, 2003; Marbán *et al.*, 2008] a los de un proyecto de software tradicional, y no se ajustan a sus particularidades.

En la investigación documental realizada, se han identificado dos métodos que permiten estimar el esfuerzo inicial para desarrollar un proyecto de Explotación de Información: uno orientado a proyectos de tamaño mediano o grande y otro orientado a proyectos pequeños (sección 2.3.1). Además, se identificó un modelo de procesos para desarrollar proyectos de Explotación de Información (sección 2.3.2), con énfasis en pequeñas y medianas empresas (PyMEs), y un conjunto de procesos de Explotación de Información utilizados durante el desarrollo en la fase de modelado (sección 2.3.3). Sin embargo, no se han encontrado métricas específicas aplicables al proceso de desarrollo de un proyecto de este tipo.

2.3.1. Estimación del Esfuerzo en Proyectos de Explotación de Información

En el trabajo de Marbán [2003] se propone un método analítico de estimación para proyectos de explotación de información el cual se denomina “Matemático Paramétrico de Estimación para Proyectos de Data Mining” (en inglés Data Mining COst MOdel, o DMCoMo).

Este método es un modelo de estimación de esfuerzo paramétrico de la familia de COCOMO II, que permite estimar los meses/hombre necesarios para desarrollar un proyecto de explotación de información, desde su concepción hasta su puesta en marcha. Para realizar la estimación, se definieron seis categorías con sus factores de costo relacionados [Marbán, 2003; Marbán *et al.*, 2008], que vinculan las características más importantes de los proyectos de Explotación de Información. Estas categorías y sus factores de costos se indican en la tabla 2.2. No obstante, en validaciones realizadas por Pytel [2011] y en [Pytel *et al.*, 2012] del método DMCoMo sobre casos reales, se determinó que el mismo era aplicable a proyectos de tamaño grande y mediano.

Categorías	Descripción	Factores de Costo
Datos	Se incluyen las características que tienen que ver con la cantidad y la calidad de los datos a tratar en un proyecto de minería de datos, también se agrupan bajo esta clasificación todas las características relativas a modelos de datos y a sistemas gestores de bases de datos en los que se pueden encontrar los datos objetivos del análisis.	▪ Cantidad de Tablas ▪ Cantidad de Tuplas de las Tablas ▪ Cantidad de Atributos de las Tablas ▪ Grado de Dispersión de los Datos ▪ Porcentaje de valores Nulos ▪ Grado de Documentación de las Fuentes de Información ▪ Grado de Integración de Datos Externos
Modelos	Se incluyen aquellas características que tienen que ver con los modelos que hay que generar y que tienen en cuenta el volumen de datos que se va a utilizar para generar los modelos, la disponibilidad de técnicas para generar los modelos y la dificultad del mismo.	▪ Cantidad de Modelos a ser Creados ▪ Tipo de Modelos a ser Creados ▪ Cantidad de Tuplas de los Modelos ▪ Cantidad y Tipo de Atributos por cada Modelo ▪ Cantidad de Técnicas Disponibles para cada Modelo
Plataforma	Agrupa las características que tienen que ver con los almacenes de datos y su localización.	▪ Cantidad y Tipo de Fuentes de Información Disponibles ▪ Distancia y Medio de Comunicación entre Servidores de Datos
Técnicas y Herramientas	Contempla las características de las técnicas y herramientas de minería de datos que se van a utilizar en el proyecto. Estas características se centran principalmente en el nivel de formación que requieren, la amigabilidad de las mismas y el número de técnicas que soportan.	▪ Herramientas disponibles para ser usadas ▪ Grado de Compatibilidad de las Herramientas con Otros Software ▪ Nivel de Formación de los Usuarios en las Herramientas
Proyecto	Incluye las características relativas a los departamentos (áreas) y localizaciones para las que se desarrolla el proyecto y acerca de la documentación que es necesaria generar durante la realización del mismo..	▪ Cantidad de Departamentos Involucrados en el Proyecto ▪ Grado de Documentación que es necesario generar ▪ Cantidad de Sitios donde se realizará el Desarrollo y su Grado de Comunicación
Equipo de Trabajo	Incluye aquellas características relacionadas con el equipo de trabajo que participa en el proyecto (dirección, implementadores, expertos, etc.). Estas características evalúan el conocimiento y la capacidad requerida para llevar a cabo cada una de las tareas del proyecto.	▪ Grado de Familiaridad con el Tipo de Problema ▪ Grado de Conocimiento de los Datos ▪ Actitud de los Directivos

Tabla 2.2. *Características del Método de Estimación de Esfuerzo DMCoMo*

Por otra parte, y teniendo en cuenta que los proyectos de Explotación de Información mayormente son requeridos por empresas PyMEs, en el trabajo de [Pytel *et al.*, 2012] se propone un método de estimación de esfuerzo orientado a las características de estas empresas, encuadrándose en el contexto de los proyectos de tamaño pequeño. Este nuevo método de estimación toma de referencia el método DMCoMo pero con una menor cantidad de factores de costo. Las categorías definidas en el método de estimación para proyectos de Explotación de Información para PyMEs y sus factores de costos se indican en la tabla 2.3.

Categorías	Descripción	Factores de Costo
Datos	Se incluyen las características de los repositorios disponibles (públicos o privados de la organización) como la tecnología con que se encuentran implementadas.	▪ Cantidad y Tipo de los Repositorios de Datos Disponibles ▪ Cantidad de Tuplas Disponibles en la Tabla Principal ▪ Cantidad de Tuplas Disponibles en Tablas Auxiliares ▪ Nivel de Conocimiento sobre los Datos
Proyecto	Incluye las características relacionadas a los objetivos de explotación de información a cumplir y al grado de compromiso que la alta gerencia y los niveles intermedios de la organización tienen con el proyecto.	▪ Tipo de Objetivo de Explotación de Información ▪ Grado de Apoyo de los Miembros de la Organización
Recursos	Contempla las características relacionadas con el equipo de trabajo que participa en el proyecto y las herramientas disponibles para la realización del mismo.	▪ Nivel de Conocimiento y Experiencia del Equipo de Trabajo ▪ Funcionalidad de las Herramientas Disponibles

Tabla 2.3. *Características del Método de Estimación de Esfuerzo para PyMEs*

2.3.2. Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información

Las etapas de desarrollo de los proyectos de Explotación de Información no coinciden naturalmente con las fases mediante las cuales se desarrollan los proyectos de software tradicionales [Vanrell, 2011], ya que estas etapas están completamente relacionadas con las distintas transformaciones que sufren los datos a lo largo del desarrollo del proyecto. En este sentido, el Modelo de Procesos para Proyectos de Explotación de Información propuesto por Vanrell [2011] plantea dos procesos principales: uno vinculado a la administración de proyectos de explotación de información y otro relacionado con el desarrollo del mismo. Para el interés de este trabajo, nos centraremos en el Modelo de Proceso de Desarrollo, cuyos subprocesos y tareas fueron definidas a partir de las fases de desarrollo planteadas por la metodología CRISP-DM [Chapman *et al.*, 2000].

En la figura 2.1 se observan los subprocesos que componen el Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información definido por Vanrell [2011], en el orden secuencial natural de los mismos.

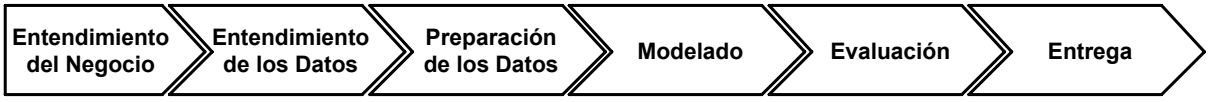

Figura 2.1. *Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información [Vanrell, 2011]*

A continuación se describen cada uno de los subprocesos definidos por Vanrell [2011]:

En el subproceso de *Entendimiento del Negocio* se deben entender los objetivos del proyecto de explotación de información y determinar los criterios de éxito a alcanzar para lograr dichos objetivos.

El subproceso *Entendimiento de los Datos* comienza con la recolección inicial de datos y las acciones para familiarizarse con ellos, se identifican los problemas de calidad que puedan presentar con los datos y los subconjuntos interesantes de datos que puedan contribuir con las primeras hipótesis de información oculta.

El subproceso de *Preparación de los Datos* cubre todas las actividades para construir el conjunto de datos final desde los datos iniciales. En este caso, se toma la información disponible para su manipulación (selección de tablas, atributos y registros), transformación (limpieza de datos, cambios de formato, construcción de atributos adicionales), y presentación (integración de los datos necesarios en una única tabla), con el objetivo de efectuar su procesamiento a través de técnicas de minería de datos. Las tareas de este subproceso pueden ser realizadas muchas veces y sin un orden preestablecido.

El subproceso *Modelado* incluye la selección de las técnicas de modelado y calibración de sus parámetros a los valores óptimos, la construcción de uno o varios modelos con la mayor calidad desde la perspectiva de análisis, el diseño de las pruebas y la evaluación del modelo generado. Suelen existir distintas técnicas para un mismo problema de explotación de información (árboles de decisión, reglas de decisión, redes neuronales, etc.) y cada una de ellas tener ciertos requisitos sobre los datos, por lo que muchas veces es necesario volver al subproceso de Preparación de los Datos.

El subproceso de *Evaluación* requiere la revisión de los pasos ejecutados para la construcción del/los modelo/s para asegurarse de lograr los objetivos de negocio. Al final de este subproceso se deber poder tomar una decisión respecto de la utilización de los resultados y obtener la aprobación de los modelos generados para el proyecto.

Por último, el subproceso de *Entrega* requiere la generación de un reporte y la presentación final del proyecto de Explotación de Información. Este reporte debe presentar los resultados de manera comprensible en orden a lograr un incremento del conocimiento.

En la tabla 2.4 se muestran las tareas y salidas asociadas a cada uno de estos subprocesos. Claramente se observa que los subprocesos y tareas indicadas difieren de las etapas definidas para un proyecto de desarrollo de software tradicional (inicio, requerimientos, análisis y diseño, construcción, integración y pruebas y cierre).

Subprocesos	Tareas	Salida
Entendimiento del Negocio	Determinar las metas del proyecto de Explotación de Información	Metas del proyecto de Explotación de Información
		Criterios de éxito del proyecto de Explotación de Información
Entendimiento de los Datos	Reunir los datos iniciales	Reporte de datos iniciales
	Describir los datos	Reporte de descripción de los datos
	Explorar los datos	Reporte de exploración de los datos
	Verificar la calidad de los datos	Reporte de calidad de los datos
Preparación de los Datos	Tareas preparatorias	Datasets
		Descripción de los Datasets
	Seleccionar los datos	Justificación de inclusión/exclusión
	Limpiar los datos	Reporte de limpieza de datos
	Construir los datos	Atributos derivados
		Registros generados
	Integrar los datos	Datos combinados (combinación de tablas y agregaciones)
	Formatear los datos	Datos formateados
Modelado	Seleccionar la técnica de modelado	Técnica de modelado
		Suposiciones de modelado
	Generar el diseño de test	Diseño de test
	Construir el modelo	Establecimiento de parámetros
		Modelos
		Descripción del modelo
	Evaluar el modelo	Evaluación del modelo
		Revisión de los parámetros establecidos
Evaluación	Evaluar resultados	Evaluación de los resultados de Explotación de Información respecto a los criterios de éxito
		Modelos aprobados
	Revisar el proceso	Revisión del proceso
	Determinar próximos pasos	Lista de posibles decisiones
		Decisiones
Entrega	Producir un reporte final	Reporte final
		Presentación final

Tabla 2.4. *Tareas vinculadas al Modelo de Proceso de Desarrollo [Vanrell, 2011]*

2.3.3. Procesos de Explotación de Información

En el trabajo realizado por Britos [2008] se definen cinco procesos de Explotación de Información que pueden ser considerados dentro de la etapa de Modelado del desarrollo de un proyecto. Los procesos de explotación de información definidos son los siguientes:

- Descubrimiento de Reglas de Comportamiento
- Descubrimiento de Grupos
- Ponderación de Interdependencia de Atributos
- Descubrimiento de Reglas de Pertenencia a Grupos
- Ponderación de Reglas de Comportamiento o de la Pertenencia a Grupos

El proceso de *Descubrimiento de Reglas de Comportamiento* se utiliza al querer identificar condiciones para obtener resultados del dominio del problema. Puede ser utilizado para descubrir las características del local más visitado por los clientes o establecer las características de los clientes con alto grado de fidelidad a la marca.

El proceso de *Descubrimiento de Grupos* es útil en los casos en que se necesita identificar una partición dentro de la información disponible en el dominio de un problema. Como ejemplos de este tipo de procesos se mencionan la identificación de tipos de llamadas que realizan los clientes de una empresa de telecomunicaciones o la identificación de grupos sociales con las mismas características, entre otros.

El proceso de *Ponderación de Interdependencia de Atributos* se utiliza cuando se desea identificar los factores con mayor incidencia sobre un determinado resultado de un problema. Son ejemplos aplicables a este proceso la determinación de factores que poseen incidencia sobre las ventas o la individualización de atributos clave que convierten en vendible a un determinado producto.

El proceso de *Descubrimiento de Reglas de Pertenencia a Grupos* es utilizado cuando se necesita identificar las condiciones de pertenencia a cada una de las clases en una partición desconocida pero que se encuentra presente en la masa de información disponible sobre el dominio del problema. Este tipo de proceso puede ser utilizado para la segmentación etaria de estudiantes y el comportamiento de cada segmento o la determinación de las clases de las llamadas telefónicas en una región y caracterización de cada clase, entre otros.

Por último, el proceso de *Ponderación de Reglas de Comportamiento de la Pertenencia a Grupos* se utiliza cuando se requiere identificar las condiciones con mayor incidencia sobre la obtención de un determinado resultado en el dominio del problema, ya sea por la mayor medida en la que inciden sobre su comportamiento o las que mejor definen la pertenencia a un grupo. Como ejemplos de este tipo de proceso se puede citar la identificación del factor dominante que incide en el alza de ventas de un producto dado o el rasgo con mayor presencia en los clientes con alto grado de fidelidad a la marca, entre otros.

Para cada uno de los procesos mencionados anteriormente, se propone la utilización de distintas tecnologías, en su mayoría provenientes del campo del aprendizaje automático [García-Martínez *et al.*, 2003]. No obstante, los procesos son independientes de la tecnología que se utilice para resolverlos.

2.4. Necesidad de Métricas en Proyectos de Explotación de Información

Los proyectos de Explotación de Información, necesitan aplicar un proceso de desarrollo y utilizar métricas para evaluar distintos aspectos del desarrollo del proyecto y el cumplimiento de los criterios de éxito del problema de negocio al finalizar el mismo. En ese sentido, el Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información propuesto por Vanrell [2011] define la manera de desarrollar en forma exitosa un proyecto de Explotación de Información. Sin embargo, no especifica qué métricas utilizar para evaluar la calidad del proceso, del producto y de los resultados obtenidos.

En [Basso *et al.*, 2013] se menciona que los subprocesos y tareas contemplados por este Modelo de Proceso podrían ser una guía para proponer un conjunto de métricas significativas aplicables al desarrollo de proyectos de Explotación de Información, con énfasis en las características de las empresas PyMEs. Asimismo, se indica que las características contempladas por los dos métodos de estimación del esfuerzo definidos para un proyecto de Explotación de Información [Marbán, 2003; Marbán *et al.*, 2008; Pytel *et al.*, 2012], podrían servir de referencia para proponer una clasificación para las métricas. Dentro de ese contexto, en [Basso *et al.*, 2013] se esbozan los primeros lineamientos y conclusiones parciales obtenidas.

Por otra parte, de la investigación documental realizada se observa que a partir de las métricas existentes de la Ingeniería de Software y la Ingeniería del Conocimiento, también podrían considerarse aquellas adaptables a los proyectos de Explotación de Información (sección 3.2).

2.4.1. Métricas Existentes aplicables a Proyectos de Explotación de Información

Desde el punto de vista de la Ingeniería de Software y la Ingeniería del Conocimiento se plantea la necesidad de analizar parámetros objetivos y prácticos de las métricas existentes, de manera de encontrar aquellas que sean aplicables al desarrollo de proyectos de Explotación de Información.

La Ingeniería del Software utiliza a los Puntos de Función [Albretch, 1979] como técnica para medir el tamaño del proyecto. Esta técnica considera la estimación empírica [Fairley, 1992] dada por la relación entre el esfuerzo requerido para construir el software y las características identificadas del mismo, tales como entradas externas (atributos/campos), archivos de interface, salidas, consultas y archivos lógicos internos (tablas). Las cantidades para cada una de estas características son ajustadas a través de la ponderación y de factores de complejidad para obtener un tamaño expresado en puntos de función.

En un proyecto de Explotación de Información, el tamaño no se mide a través de puntos de función sino que se establecen tres rangos de tamaño de proyectos (grandes, medianos y pequeños) [Pytel, 2011], los cuales se determinan a partir de las características y los valores de los factores de costo considerados por el método DMCoMo [Marbán 2003; Marbán *et al.*, 2008]. No obstante, el número de tablas y número de atributos podrían ser métricas a tener en cuenta ya que representan las fuentes de datos necesarias para desarrollar el proyecto y la disponibilidad de una cantidad suficiente de datos para aplicar explotación de información.

Por otra parte, se observa que podrían considerarse otras métricas de la Ingeniería de Software, entre las que se mencionan:

- Porcentaje de tareas completadas
- Porcentaje de tiempo total dedicado a las pruebas
- Porcentaje de error en la estimación del tiempo
- Cantidad de personas que trabajan en el proyecto
- Cantidad de personas requeridas por cada fase del proceso
- Cantidad de horas trabajadas
- Tiempo transcurrido
- Distribución del esfuerzo por cada fase del proceso
- Costo de hs/persona
- Costo del Desarrollo
- Cantidad de software desarrollado por unidad de tiempo de trabajo (productividad)

Estas métricas, sumadas a la complejidad propia del producto y del proyecto, estarían vinculadas al esfuerzo y duración real requeridos para realizar cada una de las tareas del proceso de desarrollo del proyecto de Explotación de Información, obteniéndose de esta manera las métricas que miden el progreso ó avance, el desvío de esfuerzo (estimado vs. real) y el costo real del proyecto.

Dentro del campo de la Ingeniería del Conocimiento, se considera que las siguientes métricas serían de aplicación al desarrollo de proyectos de Explotación de Información:

- Número de Atributos: esta métrica permitiría conocer si se dispone de una cantidad suficiente de datos para aplicar explotación de información, tal como se mencionó anteriormente.
- Número de Reglas: esta métrica permitiría conocer qué grado de representatividad tienen los datos utilizados en el proyecto, luego de aplicar los procesos de descubrimiento de reglas de

comportamiento o de pertenencia a grupos, en el desarrollo de los modelos de explotación de información.

- Número de Atributos de una Regla / Número de Atributos: esta métrica permitiría conocer la proporción de atributos utilizados por las reglas generadas en el modelo de explotación de información, luego de aplicar los procesos de descubrimiento de reglas de comportamiento o de pertenencia a grupos, respecto de los considerados en el desarrollo del mismo.

3. CONCLUSIONES

En este Capítulo se presenta un resumen de los resultados del trabajo de investigación (sección 3.1) y las futuras líneas de investigación surgidas a partir de él (sección 3.2).

3.1. Resumen de los Resultados del Trabajo

Al igual que los proyectos de desarrollo de software tradicionales, los proyectos de Explotación de Información necesitan métricas para medir y controlar el avance del proyecto, evaluar el esfuerzo aplicado, la calidad del proceso y los productos desarrollados. Sin embargo, y pese al interés que existe en la comunidad científica de dar solución a este tema, no se han propuesto hasta el momento métricas significativas aplicables al desarrollo de proyectos de este tipo.

En la Ingeniería de Software y la Ingeniería del Conocimiento, existen métricas e indicadores que permiten evaluar y analizar distintos atributos y características de los productos y procesos, asegurando la calidad del software producido. Se ha observado que algunas de estas métricas podrían ser aplicables al desarrollo de proyectos de Explotación de Información, adaptándolas a su contexto específico.

Por otra parte, se ha identificado que a partir de los factores de costo y características consideradas por los métodos de estimación de proyectos de Explotación de Información existentes, podría plantearse una categorización de métricas para estos proyectos. Asimismo, los subprocesos y tareas definidas en el Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información, podrían ser la referencia para una propuesta de métricas con indicadores específicos para estos proyectos que permitan evaluar la calidad del proceso y el producto obtenido.

En vista de lo expuesto anteriormente creemos que es justificada la definición de una Propuesta de Métricas para Proyectos de Explotación de Información, orientada a Pequeñas y Medianas Empresas (PyMES), utilizando como base el Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información.

3.2. Futuras Líneas de Investigación

A partir del desarrollo de este trabajo, se ha identificado la necesidad de establecer los siguientes objetivos como futuras líneas de investigación:

I. Definir una categorización de métricas que sean aplicables al proceso de desarrollo de Proyectos de Explotación de Información, a partir de las características consideradas por los métodos de estimación de esfuerzo existentes para estos proyectos.

II. Proponer un conjunto de métricas significativas para Proyectos de Explotación de Información siguiendo los lineamientos, subprocesos y tareas enunciadas en el Modelo de Proceso Desarrollo definido en [Vanrell, 2011].

III. Establecer indicadores y parámetros adecuados para las métricas propuestas, que se correspondan con las características consideradas en los proyectos de empresas PyMEs.

4. REFERENCIAS

Abraham, A. (2003). *Business Intelligence from Web Usage Mining*. Journal of Information & Knowledge Management, 2(4): pp. 375-390.

Albrecht, A. (1979). *Measuring Aplication Development Productivity*. Proc of IBM applications. Development Joint SHARE/GUIDE Symposium, Monterrey, pp. 83-92.

Basso, D., Rodríguez, D., García-Martínez, R. (2013). *Propuesta de Métricas para Proyectos de Explotación de Información*. Workshop de Bases de Datos y Minería de Datos. Proceedings XIX Congreso Argentino de Ciencias de la Computación. Pag. 983-992. ISBN 978-987-23963-1-2.

Briand L.C., Daly J.W., Wüst J. (1998). A Unified Framework for Cohesion Measurement in Object-Oriented Systems. Empirical Software Engineering, 3, pp. 65-117.

Britos, P. (2008). *Procesos de Explotación de Información basados en Sistemas Inteligentes*. Tesis Doctoral. Universidad Nacional de La Plata. Facultad de Informática. Argentina.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step Data Mining guide*. U.S.A.

Cooley, R. (2003). *The Use of Web Structure and Content to Identify Subjectively Interesting Web Usage Patterns*. ACM Transactions on Internet Technology, 3(2): pp. 93-116.

Curtis, B., Kellner, M., Over, J. (1992). *Process Modelling*. Communications of the ACM, 35(9): 75-90.

Estayno, M., Dapozo, G., Cuenca Pletsch, L., Greiner, C. (2009). *Modelos y Métricas para Evaluar la Calidad de Software*. XI Workshop de Investigadores en Ciencias de la Computación de la Red de Universidades con Carreras en Informática (RedUNCI). ISBN: 978-950-605-570-7, pp. 382-388.

Fairley, R.E. (1992) *Recent Advances in Software Estimation Techniques*. Proc. 14th Int'l Conf. Software Eng., ACM Press, New York.

Fenton, N., Pfleeger S. (1997). *Software Metrics: A Rigorous Approach*. Londres, Chapman & Hall.

Fenton N., Neil M. (1999). *Software metrics: Successes, failures and new directions*. The Journal of Systems and Software; 47(2–3):149–157.

Firestone, J. (2004). Knowledge Management Metrics Development: A Technical Approach. Published on-line by Executive Information Systems, Inc. http://www.dkms.com/papers/kmmeasurement.pdf. Último acceso Junio 2014.

Gambin D., Pallota, E. (2009). *Minería de Datos aplicada a Cultivos de Maíz*. Trabajo Final de Ingeniería en Informática. Universidad de Buenos Aires (UBA). Facultad de Ingeniería. Argentina.

García-Martínez, R., Servente, M., Pasquini, D. (2003). *Sistemas Inteligentes*. Editorial Nueva Librería. Buenos Aires.

García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo-Cattaneo, F., Rodríguez, D., Pytel, P., Vanrell. J. (2011). *Towards an Information Mining Engineering. En Software Engineering, Methods, Modeling and Teaching*. Sello Editorial Universidad de Medellín. ISBN 978-958-8692-32-6. Páginas 83-99.

García-Martínez, R., Britos, P., Rodríguez, D. (2013). *Information Mining Processes Based on Intelligent Systems*. Lecture Notes on Artificial Intelligence, 7906: 402-410. ISBN 978-3- 642-38576-6.

Hauge, O., Britos, P., García-Martínez, R. (2006). *Conceptualization Maturiry Metrics for Expert Systems*. IFIP International Federation for Information Processing, Volume 217, Artificial Intelligence in Theory and Practice, ed. M. Bramer, (Boston: Springer), pp. 435-444.

IEEE (1993). IEEE Standard Glossary of Software Engineering Terminology.

ISO/IEC 9126-1. (2001). Software engineering - Product quality - Part 1 Quality model. http://www.iso.org/iso/home.html.

ISO/IEC 9126-2. (2003). Software engineering - Product quality - Part 2 External metrics. http://www.iso.org/iso/home.html.

ISO/IEC 9126-3. (2003). Software engineering - Product quality - Part 3 Internal metrics. http://www.iso.org/iso/home.html.

ISO/IEC 9126-4. (2004). Software engineering - Product quality - Part 4 Quality in use metrics. http://www.iso.org/iso/home.html.

ISO/IEC 25010. (2011). Systems and software engineering - Software Quality Requirements and Evaluation (SQuaRE) - System and software quality models. http://www.iso.org/iso/home.html.

Kan, S. H., Parrish, J., Manlove, D. (2001). *In-process metrics for software testing*. IBM Systems Journal, 40(1): 220-241.

Larose, D. (2005). *Discovering Knowledge in Data, an introduction to Data Mining*. John Wiley & Sons. EEUU.

Maimon, O., Rokach, L. (2005). *The Data Mining and Knowledge Discovery Handbook*. Springer Science + Business Media Publishers.

Marbán Gallego, O. (2003). *Modelo Matemático Paramétrico de Estimación para Proyectos de Data Mining (DMCOMO)*. Tesis Doctoral. Departamento de Lenguajes y Sistemas e Ingeniería Software. Facultad de Informática. Universidad Politécnica de Madrid (UPM) . España.

Marbán, O., Menesalvas E., Fernández-Baizán, C. (2008). *A cost model to estimate the effort of datamining projects (DMCoMo)*. Elsevier. Science Direct. Information System. Volume 33, Issue 1 (March 2008). Pp. 133–150.

Martins, S. (2013). *Derivación del Proceso de Explotación de Información desde el Modelado del Negocio*. Trabajo Final de Licenciatura en Sistemas. Departamento de Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús (UNLa). Argentina.

McCall, J.A., Richards, P.K., Walters, G.F. (1977) – *Factors in Software Quality*. Vols. I, II, III. NTISAD-AO49-014, 015, 055.

McDermid, J. (1991). *Software Engineer's Reference Book*. Editorial Butterworth-Heinemann Ltd.

Mobasher, B., Cooley R., Srivastava J. (1999). *Creating adaptive web sites through usage-based clustering of URLs*. Proceedings Workshop on Knowledge and Data Engineering Exchange, pp. 19-25.

Negash, S., Gray, P. (2008). *Business Intelligence. In Handbook on Decision Support Systems*. 2, ed.eds. F. Burstein y C. Holsapple (Heidelberg, Springer), pp. 175-193.

Negro, P. (2008). *Umbrales para Métricas Orientadas a Objetos*. Tesis de Maestría en Tecnología Informática. Universidad Abierta Interamericana (UAI). Facultad de Tecnología Informática. Argentina.

Pfleeger, S. L. (1997). *Assessing Software Measurement*. IEEE Software March/April, pp. 25-26.

Porta García, S., Chávez Márquez, N., Labañino, Y. (2012). *Indicadores de Calidad para Software de Simulación*. Publicación en Serie Científica de la Universidad de las Ciencias Informáticas. RNPS: 2343. ISSN: 2306-2495 – Temática Calidad de Software. No. 10, Vol. 5. http://publicaciones.uci.cu/index.php/SC/article/viewFile/1004/587. Último acceso Junio 2014.

Pollo-Cattaneo, M.F. (2007). *Sistemas Expertos. Conceptualización y Métricas de Madurez*. Trabajo Final de Especialidad en Ingeniería de Sistemas Expertos. Instituto Tecnológico de Buenos Aires (ITBA). Argentina.

Pollo-Cattaneo, F. Fernández E., Merlino, H. Rodríguez, D., Britos, P., García-Martínez, R. (2008). *Métricas de Madurez en Conceptualización de Sistemas Expertos. Casos de Estudio*. VII Jornadas Iberoamericanas de Ingeniería del Software e Ingeniería del Conocimiento. Guayaquil, Ecuador. Publicación en Sección II-c pp.107-115.

Pressman, R. (2005). *Ingeniería de Software. Un enfoque práctico*. *Sexta Edición*. Parte IV. Cap. 15, 21-26. Editorial McGraw-Hill.

Pyle, D. (2003). *Business Modeling and Business intelligence*. Morgan Kaufmann Publishers.

Pytel, P. (2011). *Método de Estimación de Esfuerzo para Proyectos de Explotación de Información. Herramienta para su Validación*. Tesis de Magister en Ingeniería del Software. Universidad Politécnica de Madrid (UPM). Instituto Tecnológico de Buenos Aires (ITBA). Argentina.

Pytel, P., Britos, P., García-Martínez, R. (2012). *Comparación de Métricas de Estimación para Proyectos de Explotación de Información*. Proceedings of Latin American Congress on Requirements Engineering and Software Testing. Pág. 29-37. ISBN 978-958-46-0577-1.

Rodríguez G., González J., Dávila G. (1998): *La norma ISO 9001 en una fábrica de software a la medida*. Revista Soluciones Avanzadas, pp.27.

Sanders, J., Curran, E. (1995). *Software Quality. A Framework for Success in Software Development and Support*. Addison Wesley. Volume 5, Issue 4. ISBN: 0-201-631989.

SAS. (2008). *SAS Enterprise Miner: SEMMA*. http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html. Último acceso Junio 2014.

Schiefer, J., Jeng, J., Kapoor, S., Chowdhary, P. (2004). *Process Information Factory: A Data Management Approach for Enhancing Business Process Intelligence*. Proceedings 2004. IEEE International Conference on ECommerce Technology. Pág. 162-169.

Screpnik, C. (2013). *Métricas Aplicables a la Evaluación de Sitios e-government y su Impacto Social*. Tesis de Especialidad en Ingeniería de Software. Universidad Nacional de La Plata. Facultad de Informática. Argentina.

SEI (2010). *CMMI® for Development, Version 1.3*. Carnegie Mellon University, Software Engineering Institute. http://resources.sei.cmu.edu/asset_files/TechnicalReport/2010_005_001_15287.pdf. Último acceso Junio 2014.

Srivastava, J., Cooley, R., Deshpande, M., Tan, P. (2000). *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*. SIGKDD Explorations, 1(2): pp.12-23.

Stefanovic, N., Majstorovic. V., Stefanovic, D. (2006). *Supply Chain Business Intelligence Model*. Proceedings 13th International Conference on Life Cycle Engineering. Pág. 613-618.

Umapathy, K. (2007). *Towards Co-Design of Business Processes and Information Systems Using Web Services*. Proceedings 40th Annual Hawaii International Conference on System Sciences. Pág. 172-181.

Vanrell, J. A., Bertone, R., García-Martínez, R. (2010a). *Modelo de Proceso de Operación para Proyectos de Explotación de Información*. Anales del XVI Congreso Argentino de Ciencias de la Computación. Pág. 674-682. ISBN 978-950-9474-49-9.

Vanrell, J. (2011). *Un Modelo de Procesos para Proyectos de Explotación de Información*. Tesis de Maestría en Ingeniería en Sistemas de Información. Escuela de Posgrado. Universidad Tecnológica Nacional. Facultad Regional Buenos Aires.