Identificación de Tareas Críticas en una Metodología de Desarrollo de Proyectos de Explotación

Pytel, P., Pollo-Cattaneo, F., Rodríguez, D., Britos, P., García-Martínez, R.

Grupo Investigación en Sistemas de Información. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús.
Grupo de Investigación en Explotación de Información. Sede Andina (El Bolsón). Universidad Nacional de Río Negro.
Grupo de Estudio en Metodologías de Ingeniería de Software. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional.
ppytel@gmail.com, fpollo@posgrado.frba.utn.edu.ar, drodrigu@unla.edu.ar, paobritos@gmail.com, rgarcia@unla.edu.ar

Resumen. Un proceso de explotación de información puede definirse como un conjunto de tareas relacionadas lógicamente, que se ejecutan para extraer conocimiento no-trivial que reside de manera implícita en los datos disponibles en distintas fuentes de información. Una metodología de explotación de información permite gestionar la complejidad de estos procesos de manera uniforme. Entre estas metodologías, la comunidad científica considera probada a la metodología CRISP-DM. Se considera necesario identificar cuales son las tareas críticas de esta metodología que deben ser consideradas para realizar una planificación exitosa. En este artículo, se introduce brevemente la metodología CRISP con énfasis en las tareas asociadas a sus subfases, se presentan resultados experimentales que muestran el resultado del proceso de descubrimiento de reglas de pertenencia a grupos a la información de proyectos de explotación de información para pequeños y medianos emprendimientos.

Palabras Clave. Proceso de explotación de información. Metodología de explotación de información. CRISP-DM. Planificación de proyectos de explotación de información. Minería de Datos.

1. Introducción

La Explotación de Información consiste en la extracción de conocimiento no-trivial que reside de manera implícita en los datos disponibles en distintas fuentes de información [1]. Dicho conocimiento es previamente desconocido y puede resultar útil para algún proceso [2]. Para un experto, o para el responsable de un sistema de información, normalmente no son los datos en sí lo más relevante, sino el conocimiento que se encierra en sus relaciones, fluctuaciones y dependencias. Con Explotación de Información se aborda la solución a problemas de predicción, clasificación y segmentación [3].

Por otro lado, la Ingeniería en Software utiliza diferentes modelos y metodologías para obtener proyectos de informática con gran nivel de previsibilidad y calidad. Estos permiten controlar la calidad final de producto a desarrollar estableciendo controles sobre cada una de las etapas que intervienen en el proceso productivo, entendiendo por proceso productivo no sólo a la producción en sí misma, sino también a las tareas relacionadas a la gestión de un proyecto y de la empresa que lo desarrolla [4]. Una metodología de Explotación de Información involucra, en general las siguientes fases [5]: 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. Para el desarrollo de proyectos de explotación de información entre las cuales se destacan CRISP [6], P_3TQ [7] y SEMMA [8].
Se ha seleccionado la metodología CRISP-DM como metodología de referencia dado que, cuando se comparan las tres metodologías mencionadas anteriormente, la comunidad científica considera que esta última contiene más elementos a nivel operación de las otras [9]. En este contexto, se realiza una introducción de CRISP-DM (sección 2); luego se describe el problema detectado en el ámbito de los proyectos de explotación (sección 3); se presenta el experimento realizado (sección 4) y se discuten los resultados obtenidos (sección 5) para finalmente presentar algunas conclusiones preliminares (sección 6).

2. Metodología de Desarrollo de Proyectos de Explotación de Información CRISP-DM

La metodología CRISP-DM [6] consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos (ver Figura 1).

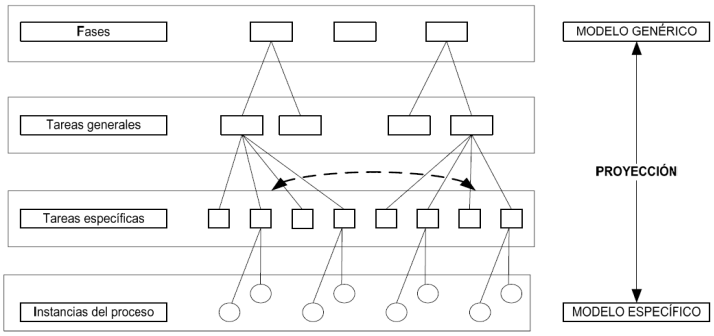

Figura 1. Esquema de los cuatro niveles de abstracción de la metodología CRISP-DM

A nivel más general, el proceso está organizado en seis fases (ver Figura 2), estando cada fase a su vez estructurada en varias tareas generales de segundo nivel o subfases. Las tareas generales se proyectan a tareas específicas, donde se describen las acciones

que deben ser desarrolladas para situaciones específicas. Así, si en el segundo nivel se tiene la tarea general "limpieza de datos", en el tercer nivel se dicen las tareas que tienen que desarrollarse para un caso específico, como por ejemplo, "limpieza de datos numéricos", o "limpieza de datos categóricos". El cuarto nivel, recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de Explotación de Información específico.

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Explotación de Información en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto (ver Figura 2). Las flechas indican las relaciones más habituales entre las fases, aunque se pueden establecer relaciones entre cualquier fase. El círculo exterior simboliza la naturaleza cíclica del proceso de modelado. En la tabla 1, se detallan las fases que componen la metodología CRISP-DM y se detalla cómo se componen cada una de ellas.

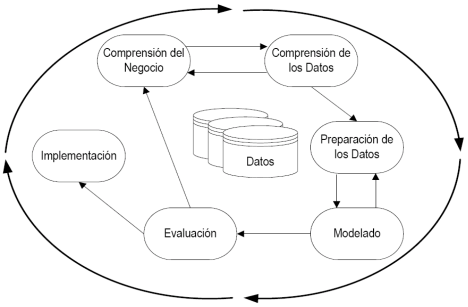

Figura 2. Fases del proceso de modelado metodología CRISP-DM.

La primera fase de análisis del problema, incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación. La segunda fase de análisis de datos comprende la recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis. Una vez realizado el análisis de datos, la metodología establece que se proceda a la preparación de los datos, de tal forma que puedan ser tratados por las técnicas de modelado. La preparación de datos (fase 3) incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. La fase de preparación de los datos, se encuentra muy relacionada con la fase de modelado, puesto que en función de la técnica de modelado que vaya a ser utilizada los datos necesitan ser procesados en diferentes formas. Por lo tanto las fases de preparación y modelado interactúan de forma sistemática. En la fase de modelado (fase 4) se seleccionan las técnicas de modelado más apropiadas para el proyecto de Explotación de Información específico. Las técnicas a utilizar en esta fase se seleccionan en función de los siguientes criterios: ser apropiada al problema, disponer de datos adecuados, cumplir los requerimientos del problema, tiempo necesario para obtener un modelo y conocimiento de la técnica.

Tabla 1. Tareas de cada fase de la metodología CRISP-DM

FASE	SUBFASES	TAREAS ASOCIADAS
1. COMPRENSIÓN DEL NEGOCIO	1.1 Determinar los objetivos de negocio	1.1.1 Antecedentes 1.1.2 Objetivos de negocio 1.1.3 Criterios de éxito del negocio
	1.2 Evaluar la situación	1.2.1 Evaluar la situación 1.2.2 Requisitos, supuestos y limitaciones 1.2.3 Riesgos y contingencias 1.2.4 Terminología 1.2.5 Costos y beneficios
	1.3 Determinar objetivos explotación de información	1.3.1 Objetivos de explotación de información 1.3.2 Criterios de éxito de la de explotación de información
	1.4 Producir el plan del proyecto	1.4.1 Plan del proyecto 1.4.2 Evaluación inicial de herramientas y técnicas
2. ENTENDIMIENTO DE LOS DATOS	2.1 Recolección inicial de datos	2.1.1 Informe inicial de recopilación de datos
	2.2 Descripción de los datos	2.2.1 Informe de descripción de datos
	2.3 Exploración de los datos	2.3.1 Informe de exploración de datos
	2.4 Verificación de calidad de los datos	2.4.1 Informe de calidad de datos
3. PREPARACION DE DATOS	3.0 Tareas preparatorias	3.0.1 Conjunto de datos 3.0.2 Descripción del conjunto de datos
	3.1 Selección de datos	3.1.1 Justificación de la inclusión / exclusión
	3.2 Limpieza de datos	3.2.1 Informe de limpieza de datos
	3.3 Construcción de datos	3.3.1 Atributos derivados 3.3.2 Registros generados
	3.4 Integración de los datos	3.4.1 Datos combinados
	3.5 Formato de datos	3.5.1 Datos reformateadas
4. MODELADO	4.1 Selección de la técnica de modelado	4.1.1 Técnica de Modelado 4.1.2 Supuestos del modelado
	4.2 Generación del diseño del ensayo	4.2.1 Prueba de diseño
	4.3 Construcción del modelo	4.3.1 Configuración de parámetros 4.3.2 Modelos 4.3.3 Descripción del modelo
	4.4 Evaluación del modelo	4.4.1 Evaluación del modelo 4.4.2 Revisión de la configuración de parámetros
5. EVALUACIÓN	5.1 Evaluar los resultados	5.1.1 Evaluación de los resultados de la explotación de información con respecto a los criterios de éxito del negocio 5.1.2 Modelos aprobados
	5.2 Proceso de revisión	5.2.1 Revisión del proceso
	5.3 Determinación de los próximos pasos	5.3.1 Lista de posibles acciones 5.3.2 Decisión
6. IMPLANTACION	6.1 Plan de implantación	6.1.1 Ejecución del plan de implantación
	6.2 Plan de vigilancia y mantenimiento	6.2.1 Ejecución del plan de monitoreo y mantenimiento
	6.3 Producción final	6.3.1 Informe final 6.3.2 Presentación final
	6.4 Revisión del proyecto	6.4.1 Documentación de la experiencia

Antes de proceder al modelado de los datos se debe de establecer un diseño del método de evaluación de los modelos, que permita establecer el grado de bondad de los modelos. Una vez realizadas estas tareas genéricas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos. En la fase de evaluación (fase 5), se evalúa el modelo, no desde el punto de vista de los datos, sino desde el cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la implementación del modelo de explotación (fase 6).

3. El Problema

La gestión de un proyecto de software comienza con un conjunto de actividades que se denominan planificación del proyecto. Antes de que el proyecto comience, se debe realizar una estimación: del trabajo a ejecutar, de los recursos necesarios y del tiempo que transcurrirá desde el comienzo hasta el final de su realización [10]. Los proyectos de explotación de información también requieren de un proceso de planificación, sin embargo, dada las diferencias que existente entre un proyecto convencional de construcción de software y un proyecto de explotación de información, los métodos usuales de estimación no son aplicables [11]. De esta manera, se considera necesario identificar cuales son las tareas críticas de la metodología para desarrollo para proyectos de explotación de información. Al conocer las tareas críticas será posible gestionarlas cuidadosamente durante el desarrollo del proyecto y así intentar reducir posibles problemas que se puedan presentar.

4. El Experimento

El objetivo del experimento es identificar las tareas críticas de la metodología CRISP-DM a partir de los tiempos que insumen la ejecución de cada una de las tareas asociadas a las subfases de la Metodología CRISP-DM con foco en proyectos para pequeños y medianos emprendimientos.
Para ello se aplica el proceso de descubrimiento de reglas de pertenencia a grupos [12]. Este proceso (ver Figura 3) permite identificar cuáles son las condiciones de pertenencia a cada una de las clases en una partición desconocida “a priori”, pero presente en la masa de información disponible sobre el dominio de problema. Para el descubrimiento de reglas de pertenencia a grupos se propone la utilización de mapas auto-organizados (SOM) [13] para el hallazgo de los mismos. Los mapas auto-organizados pueden ser aplicados a la construcción de particiones de grandes masas de información por tener la ventaja de ser tolerantes al ruido y la capacidad de extender la generalización al momento de necesitar manipular datos nuevos. Una vez identificados los grupos, y la utilización de algoritmos de inducción (Top Down Induction Decision Trees o TDIDT) [14] para establecer las reglas de pertenencia a cada uno.

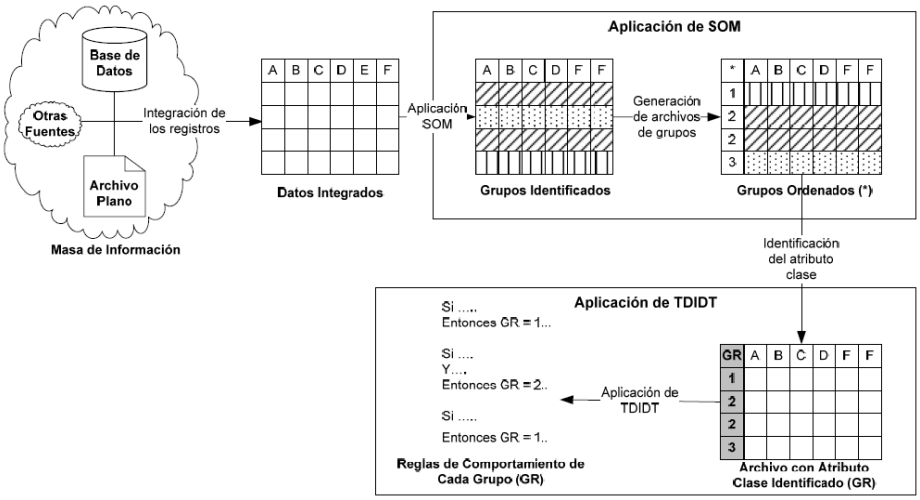

Figura 3. Esquema del proceso para descubrimiento de reglas de pertenencia a grupos

Así, los pasos seguidos en el experimento se identifican en la tabla 2.

Tabla 2. Pasos experimentales seguidos para obtener los tiempos experimentales

PASOS EXPERIMENTALES	
Paso 1:	Entrenar a estudiantes avanzados (5º año de Ingeniería en Sistemas de Información) en procesos de explotación de información y la metodología CRISP-DM. Dividir en grupos de trabajo con capacidades homogéneas.
Paso 2:	Identificar N proyectos de explotación de información para pequeños y medianos emprendimientos.
Paso 3:	Asignar cada uno de los proyectos de explotación de información a un grupo de los formados en el paso 1.
Paso 4:	Desarrollo por cada grupo del proyecto de explotación de información con registro del tiempo utilizado para desarrollar cada tarea de la metodología CRISP-DM.
Paso 5:	Integración de los tiempos obtenidos por cada grupo en un archivo único donde se dispone por cada columna las tareas y por cada fila los proyectos realizados por cada grupo con el tiempo que corresponde.
Paso 6:	Se aplica la técnica de SOM a los datos integrados obtenidos en el paso 5. Como resultado se obtiene una partición en distintos grupos (a los que se llama grupos identificados) para luego ser exportados en archivos (denominados grupos ordenados).
Paso 7:	Se aplica el algoritmo de inducción de la familia TDIDT al atributo clase de cada grupo ordenado obtenido del paso anterior (atributo “GR”) para obtener un conjunto de reglas que definen el comportamiento de cada grupo.
Paso 8:	Se interpretan las reglas obtenidas en el paso 7 para analizar el comportamiento de la duración de las tareas de la metodología CRISP-DM.

5. Interpretación de los Resultados

Como resultado de la realización de los pasos correspondientes a la recolección y preparación de los datos (pasos 1 a 5) se presenta en la tabla 3 por cada subfase de la metodología CRISP-DM el tiempo total empleado, el tiempo promedio y el % de tiempo correspondiente a la fase para los 19 proyectos de explotación de información suministrada por los alumnos.

Tabla 3. Resumen de los datos obtenidos para cada sub-fase de la Metodología CRISP-DM

FASE / SUBFASE	TOTAL DE TIEMPO (en minutos)	PROMEDIO DE TIEMPO (en minutos)	% DE TIEMPO
Fase 1 - COMPRENSIÓN DEL NEGOCIO			
Subfase 1.1 Determinar los objetivos de negocio	1.992	313	33,52%
Subfase 1.2 Evaluar la situación	1.976	105	33,25%
Subfase 1.3 Determinar objetivos explotación de información	1.039	104	17,49%
Subfase 1.4 Producir el plan del proyecto	935	55	15,74%
Fase 2 - ENTENDIMIENTO DE LOS DATOS			
Subfase 2.1 Recolección inicial de datos	1.265	211	31,49%
Subfase 2.2 Descripción de los datos	975	67	24,27%
Subfase 2.3 Exploración de los datos	1.002	51	24,94%
Subfase 2.4 Verificación de calidad de los datos	775	53	19,29%
Fase 3 - PREPARACION DE DATOS			
Subfase 3.0 Tareas preparatorias	1.333	270	26,02%
Subfase 3.1 Selección de datos	817	70	15,95%
Subfase 3.2 Limpieza de datos	700	43	13,66%
Subfase 3.3 Construcción de datos	881	37	17,20%
Subfase 3.4 Integración de los datos	457	46	8,92%
Subfase 3.5 Formato de datos	935	24	18,25%
Fase 4 - MODELADO			
Subfase 4.1 Selección de la técnica de modelado	1.142	381	15,78%
Subfase 4.2 Generación del diseño del ensayo	721	60	9,96%
Subfase 4.3 Construcción del modelo	3.879	38	53,60%
Subfase 4.4 Evaluación del modelo	1.495	204	20,66%
Fase 5 - EVALUACIÓN			
Subfase 5.1 Evaluar los resultados	1.086	123	46,35%
Subfase 5.2 Proceso de revisión	480	57	20,49%
Subfase 5.3 Determinación de los próximos pasos	777	25	33,16%
Fase 6 - IMPLANTACION			
Subfase 6.1 Plan de implantación	530	144	19,35%
Subfase 6.2 Plan de vigilancia y mantenimiento	375	28	13,69%
Subfase 6.3 Producción final	1.406	20	51,33%
Subfase 6.4 Revisión del proyecto	428	74	15,63%

Como resultado de aplicar el proceso para descubrimiento de reglas de pertenencia a grupos se obtienen las reglas indicadas en la tabla 4 que se presentan junto con su interpretación.

Tabla 4. Reglas obtenidas por el proceso para descubrimiento de reglas de pertenencia a grupos

#	REGLAS	INTERPRETACIÓN DE LA REGLA
1	SI [t1.3.2] < 7,50 Y [t3.1.1] >= 17,50 ENTONCES GR = grupo_1_2	El grupo 1.2 comprende a los proyectos donde la tarea 1.3.2 “Criterios de éxito de la de explotación de información” posee una duración menor a 7,5 minutos y la tarea 3.1.1 “Justificación de la inclusión / exclusión” posee una duración mayor o igual a 17,5 minutos.
2	SI [t1.1.1] < 20,00 Y [t3.4.1] < 7,50 Y [t3.1.1] < 17,50 ENTONCES GR = grupo_1_4	El grupo 1.4 comprende a los proyectos donde la tarea 1.1.1 “Antecedentes” posee una duración menor a 20 minutos, la duración de la tarea 3.4.1 “Datos combinados” es menor a 7,5 minutos y la de 3.1.1 “Justificación de la inclusión / exclusión” es menor a 17,5 minutos.
3	SI [t1.1.1] >= 20,00 Y [t3.4.1] < 7,50 Y [t3.1.1] < 17,50 ENTONCES GR = grupo_3_3	El grupo 3.3 comprende a los proyectos donde la tarea 1.1.1 “Antecedentes” posee una duración mayor o igual a 20 minutos, la duración de la tarea 3.4.1. “Datos combinados” es menor a 7,5 minutos y la de 3.1.1 “Justificación de la inclusión / exclusión” es menor a 17,5 minutos.
4	SI [t1.1.1] >= 27,50 Y [t3.4.1] >= 7,50 Y [t3.1.1] < 17,50 ENTONCES GR = grupo_3_1	El grupo 3.1 comprende a los proyectos donde la tarea 1.1.1 “Antecedentes” posee una duración mayor o igual a 27,5 minutos, la duración de la tarea 3.4.1. “Datos combinados” es mayor o igual a 7,5 minutos y la de 3.1.1 “Justificación de la inclusión / exclusión” es menor a 17,5 minutos.
5	SI [t1.1.2] < 17,50 Y [t1.1.1] < 27,50 Y [t3.4.1] >= 7,50 Y [t3.1.1] < 17,50 ENTONCES GR = grupo_2_1	El grupo 2.1 comprende a los proyectos donde la tarea 1.1.2 “Objetivos de negocio” posee una duración menor a 17,5 minutos, la duración de la tarea 1.1.1 “Antecedentes” es menor a 27,5 minutos, la de 3.4.1. “Datos combinados” es mayor o igual a 7,5 minutos y la de 3.1.1 “Justificación de la inclusión / exclusión” es menor a 17,5 minutos.
6	SI [t1.1.2] >= 17,50 Y [t1.1.1] < 27,50 Y [t3.4.1] >= 7,50 Y [t3.1.1] < 17,50 ENTONCES GR = grupo_4_2	El grupo 4.2 comprende a los proyectos donde la tarea 1.1.2 “Objetivos de negocio” posee una duración mayor o igual a 17,5 minutos, la duración de la tarea 1.1.1 “Antecedentes” es menor a 27,5 minutos, la de 3.4.1 “Datos combinados” es mayor o igual a 7,5 minutos y la de 3.1.1 “Justificación de la inclusión / exclusión” es menor a 17,5 minutos.
7	SI [t4.3.1] < 12,50 Y [t1.3.2] >= 7,50 Y [t3.1.1] >= 17,50 ENTONCES GR = grupo_3_1	El grupo 3.1 comprende a los proyectos donde la tarea 4.3.1 “Configuración de parámetros” posee una duración menor a 12,5 minutos, la duración de la tarea 1.3.2 “Criterios de éxito de la de explotación de información” es mayor o igual a 7,5 minutos y la de 3.1.1 “Justificación de la inclusión / exclusión” es mayor o igual a 17,5 minutos.
8	SI [t3.5.1] >= 115,00 Y [t4.3.1] >= 12,50 Y [t1.3.2] >= 7,50 Y [t3.1.1] >= 17,50 ENTONCES GR = grupo_1_3	El grupo 1.3 comprende a los proyectos donde la tarea 3.5.1 “Datos reformateadas” posee una duración mayor o igual a 115 minutos, la duración de la tarea 4.3.1 “Configuración de parámetros” es mayor o igual a 12,5 minutos, la de 1.3.2 “Criterios de éxito de la de explotación de información” es mayor o igual a 7,5 minutos y la de 3.1.1 “Justificación de la inclusión / exclusión” es mayor o igual a 17,5 minutos.
9	SI [t2.4.1] < 7,50 Y [t3.5.1] < 115,00 Y [t4.3.1] >= 12,50 Y [t1.3.2] >= 7,50 Y [t3.1.1] >= 17,50 ENTONCES GR = grupo_2_3	El grupo 2.3 comprende a los proyectos donde la tarea 2.4.1 “Informe de calidad de datos” posee una duración menor a 7,5 minutos, la duración de la tarea 3.5.1 “Datos reformateadas” es menor a 115 minutos, la de 4.3.1. “Configuración de parámetros” es mayor o igual a 12,5 minutos, la de 1.3.2 “Criterios de éxito de la de explotación de información” es mayor o igual a 7,5 minutos y finalmente la duración de la tarea 3.1.1 “Justificación de la inclusión / exclusión” es mayor o igual a 17,5 minutos.
10	SI [t2.4.1] >= 7,50 Y [t3.5.1] < 115,00 Y [t4.3.1] >= 12,50 Y [t1.3.2] >= 7,50 Y [t3.1.1] >= 17,50 ENTONCES GR = grupo_2_2	El grupo 2.2 comprende a los proyectos donde la tarea 2.4.1 “Informe de calidad de datos” posee una duración mayor o igual 7,5 minutos, la duración de la tarea 3.5.1 “Datos reformateadas” es menor a 115 minutos, la de 4.3.1 “Configuración de parámetros” es mayor o igual a 12,5 minutos, la de 1.3.2 “Criterios de éxito de la de explotación de información” es mayor o igual a 7,5 minutos y finalmente la duración de 3.1.1 “Justificación de la inclusión / exclusión” es mayor o igual a 17,5 minutos.

Así las reglas obtenidas se puede observar que los grupos están formados por las siguientes tareas:

- Tarea 3.1.1 “Justificación de la inclusión / exclusión”: 10 casos sobre 10 reglas.
- Tarea 3.4.1 “Datos combinados”: 5 casos sobre 10 reglas.
- Tarea 1.1.1 “Antecedentes”: 5 casos sobre 10 reglas.
- Tarea 1.3.2 “Criterios de éxito de la de explotación de información”: 5 casos sobre 10 reglas.
- Tarea 4.3.1 “Configuración de parámetros”: 4 casos sobre 10 reglas.
- Tarea 3.5.1 “Datos reformateadas”: 3 casos sobre 10 reglas.
- Tarea 1.1.2 “Objetivos de negocio”: 2 casos sobre 10 reglas.
- Tarea 2.4.1 “Informe de calidad de datos”: 2 casos sobre 10 reglas.

Esto nos indica que un factor determinante en la formación de grupos es la justificación de la inclusión o exclusión de los datos, seguido por la formación de datos combinados e integrados (ambas tareas involucradas con la fase de preparación de datos) y por las actividades correspondientes al entendimiento del negocio (definir antecedentes y determinar criterios del proyecto). Luego existe una única tarea correspondiente al modelado (configuración de parámetros del modelo) que aparece en menos de la mitad de las reglas obtenidas. Finalmente en sólo dos reglas se incluye una tarea de entendimiento de los datos (informe de calidad de los datos).

6. Conclusiones

Como conclusión preliminar de la interpretación de las reglas obtenidas del proceso para descubrimiento de reglas de pertenencia a grupos se puede observar el peso que tiene en el desarrollo de proyecto de explotación de información ciertas tareas de entendimiento del negocio y preparación de datos. Las tareas correspondientes a modelado y entendimiento de los datos también se consideran pero con menor importancia, pero en ningún caso se observan como determinante para formación de grupos las tareas de evaluación e implantación.
Como futura línea de investigación queda volver a realizar este experimento con datos de más proyectos para identificar si las conclusiones mencionadas se mantienen o se identifican otras tareas críticas de la metodología CRISP-DM.

7. Agradecimientos

Deseamos agradecer por su colaboración en el proyecto a los alumnos que en el año 2010 cursaron la materia Tecnologías de Explotación de Información de la carrera de Ingeniería en Sistemas de Información de la UTN FRBA.

8. Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el Proyecto de Investigación 33A105 del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús, por el Proyecto de Investigación 40B065 de la Universidad Nacional de Rio Negro - Sede Andina (El Bolsón), y por el Proyecto 25C126 de la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional.

9. Referencias

1. 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 E-Commerce Technology. Pág. 162-169.
2. Stefanovic, N., Majstorovic. V., Stefanovic, D. (2006). Supply Chain Business Intelligence Model. Proceedings 13th International Conference on Life Cycle Engineering. Pág. 613-618.
3. Umapathy, K. (2007). Towards Co-Design of Business Processes Y Information Systems Using Web Services. Proceedings 40th Annual Hawaii International Conference on System Sciences. Pág. 172-181.
4. Vanrell, J., Bertone, R., García-Martínez, R. (2010). *Modelo de Proceso de Ope-ración para Proyectos de Explotación de Información.* Anales del XVI Congre-so Argentino de Ciencias de la Computación. Pp. 674-682. ISBN 978-950-9474-49-9.
5. Maimon, O., Rokach, L. (2005). *The Data Mining Y Knowledge Discovery Handbook.* Springer Science + Business Media Publishers.
6. Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. 2000. *CRISP-DM 1.0 Step by step BIguide. Edited by SPSS.* http://www.crisp-dm.org/CRISPWP-0800.pdf. Ultimo acceso Abril 2011.
7. Pyle, D. (2003). *Business Modeling and Business intelligence*. Morgan Kaufmann Publishers.
8. SAS, (2008). *SAS Enterprise Miner: SEMMA*. http://www.sas.com/technologies/analytics /datamining/miner/semma.html. Ultimo acceso Abril 2011.
9. Vanrell, J., Bertone, R., García-Martínez, R. (2010). *Un Modelo de Procesos de Explotación de Información*. Proceedings XII Workshop de Investigadores en Ciencias de la Computación. Pp. 167-171. ISBN 978-950-34-0652-6.
10. Pressman, R. 2004. *Software Engineering: A Practitioner's Approach*. Editorial Mc Graw Hill.
11. Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. (2010). *Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información*. Anales del XVI Congreso Argentino de Ciencias de la Computación. Pág. 664-673. ISBN 978-950-9474-49-9.
12. Britos, P., García-Martínez, R. 2009. *Propuesta de Procesos de Explotación de Información*. Proceedings XV CACIC. Workshop de Base de Datos y Minería de Datos. Págs. 1041-1050. ISBN 978-897-24068-4-1.
13. Kohonen, T. (1995). *Self-Organizing Maps.* Springer Verlag Publishers.
14. Quinlan, J. (1990). *Learning Logic Definitions from Relations*. Machine Learning, 5:239-266.