VIABILIDAD Y ESTIMACIÓN DE PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

Tesista

M. Ing. Pablo PYTEL

Directores

Dr. Ramón GARCÍA MARTÍNEZ (UNLa-UNLP),

Dra. Paola BRITOS (UNRN) y Dr. Alejandro HOSSIAN (UTN-FRN)

TESIS PRESENTADA PARA OBTENER EL GRADO
DE
DOCTOR EN CIENCIAS INFORMÁTICAS

FACULTAD DE INFORMÁTICA

UNIVERSIDAD NACIONAL DE LA PLATA

AGOSTO, 2014

RESUMEN

Los proyectos de Explotación de Información son un tipo especial de proyecto de Ingeniería en Software que poseen problemas similares al existir cuestiones de gestión que pueden provocar su fracaso. Entre estas cuestiones se destaca la necesidad de establecer si es posible realizar el proyecto, si la solución es la correcta y si es posible cumplir con todas las metas y expectativas del proyecto; así como la cantidad de personas que van a participar en el equipo de trabajo durante el desarrollo del proyecto, y por cuanto tiempo. En este contexto, el presente trabajo de tesis tiene como objetivo proponer, estudiar y validar dos modelos a ser utilizados al inicio de un proyecto de Explotación de Información en el ámbito de las Pequeñas y Medianas Empresas. El primer modelo permite realizar la Evaluación de la Viabilidad del proyecto, mientras que el segundo modelo tiene como objetivo realizar la Estimación del Esfuerzo (medido en cantidad de personas por tiempo) que se requiere para llevar a cabo el proyecto de manera completa.

ABSTRACT

Information Mining projects are a special type of Software Engineering project that have similar management issues which can cause the failure of the project. Within these issues it is highlighted the necessity to establish whether it is possible to carry out the project, if the solution is correct and if it is possible to meet all the goals and expectations of the project; as long as, and the quantity of people who will participate in the team during the project, and for how long. In this context, this PhD thesis propose, study and validate two models that can be used at the beginning of an Information Mining project in the field of Small and Medium Enterprises. The first model allows the Evaluation of the Project Feasibility; while the second model calculates the Effort Estimation (measured in men-months) required to carry out the project completely.

DEDICATORIA

A mis viejos y hermanos, por estar y bancarme siempre...

AGRADECIMIENTOS

A la Facultad de Informática de la Universidad Nacional de la Plata por acogerme con generosidad de “alma mater” para que pudiera llevar a cabo mis estudios de Doctorado en Ciencias Informáticas.

Al Grupo de Investigación en Sistemas de Información del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús por recibirme para realizar la pasantía de investigación y desarrollo, y apoyarme en todas las instancias del proceso.

A mis directores de tesis: Dr. Ramón García-Martínez, Dra. Paola Britos y Dr. Alejandro Hossian por su gran guía y asistencia en la elaboración de este trabajo de tesis.

A los investigadores del Grupo de Investigación en Sistemas de Información del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús, los investigadores del Grupo de Estudio en Metodologías de Ingeniería de Software de la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional, y los investigadores del Grupo de Investigación en Explotación de Información en el Laboratorio de Informática Aplicada de la Universidad Nacional de Río Negro por el aporte de los datos utilizados en este trabajo de tesis.

A la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional por permitir formarme como docente desde hace más de 10 años.

A la M. Ing. María Florencia Pollo-Cattaneo, la Ing. Cinthia Vegega y el Ing. Hugo Ramón por su apoyo durante la realización de este trabajo de tesis.

A mis alumnos de la Facultad Regional Concepción del Uruguay de la Universidad Tecnológica Nacional que me han ayudado a definir varias de las ideas presentadas en este trabajo de tesis.

A mis amigos y compañeros de trabajo quienes siempre intentaron apoyarme para la consecución de este logro académico.

A todos los que me ayudaron a llegar aquí…

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

NOMENCLATURA

A1	Característica del Modelo de Viabilidad perteneciente a la dimensión Adecuación que evalúa si los datos disponibles se encuentran en formato digital.
A2	Característica del Modelo de Viabilidad perteneciente a la dimensión Adecuación que evalúa la cantidad de los datos disponibles.
A3	Característica del Modelo de Viabilidad perteneciente a la dimensión Adecuación que evalúa la confianza sobre los datos disponibles.
A4	Característica del Modelo de Viabilidad perteneciente a la dimensión Adecuación que evalúa si el problema de negocio puede ser resuelto por técnicas tradicionales.
A5	Característica del Modelo de Viabilidad perteneciente a la dimensión Adecuación que evalúa la estabilidad del problema de negocio.
AEXP	Coeficiente del método empírico de estimación de esfuerzo: Coeficiente de Ajuste por Experiencia del Equipo.
AREP	Factor de costo del modelo de estimación de esfuerzo: Cantidad y Tipo de los Repositorios de Datos disponibles.
CDAT	Coeficiente del método empírico de estimación de esfuerzo: Coeficiente de Complejidad de los Datos.
CMOD	Coeficiente del método empírico de estimación de esfuerzo: Coeficiente de Complejidad del Modelado.
CNEG	Coeficiente del método empírico de estimación de esfuerzo: Coeficiente de Complejidad del Negocio.
DIKW hierarchy	Acrónimo en inglés para la jerarquía de Dato-Información-Conocimiento-Sabiduría (Data-Information-Knowledge-Wisdom), también conocida como la Pirámide de la Información.
DMCoMo	Modelo Matemático Paramétrico de Estimación para Proyectos de Data Mining (en inglés, Data Mining Cost Model).
E1	Característica del Modelo de Viabilidad perteneciente a la dimensión Éxito que evalúa la tecnología en que se encuentran los datos disponibles.
E2	Característica del Modelo de Viabilidad perteneciente a la dimensión Éxito que evalúa el apoyo de los interesados de la organización al proyecto.
E3	Característica del Modelo de Viabilidad perteneciente a la dimensión Éxito que evalúa si la planificación del proyecto puede considerar buenas prácticas ingenieriles.
E4	Característica del Modelo de Viabilidad perteneciente a la dimensión Éxito que evalúa el nivel de experiencia del equipo de trabajo.
EdI	Explotación de Información.
EV	Valor Global de la Viabilidad del Proyecto.
I_d	Intervalo (formado por cuatro números) que representa el Valor de la Dimensión *d* para el Modelo de Viabilidad.
INCO	Ingeniería del Conocimiento.

KDD	Proceso de Descubrimiento de Conocimiento en Bases de Datos (en inglés, Knowledge Discovery in Databases).
KEXT	Factor de costo del modelo de estimación de esfuerzo: Nivel de Conocimiento y Experiencia del Equipo de Trabajo.
KLDS	Factor de costo del modelo de estimación de esfuerzo: Nivel de Conocimiento sobre los Datos.
LECO	Factor de costo del modelo de estimación de esfuerzo: Grado de Apoyo de los Miembros de la Organización.
MM23	Fórmula del modelo DMCoMo que utiliza los 23 factores de costo.
MM8	Fórmula del modelo DMCoMo que utiliza sólo 8 factores de costo.
OBTY	Factor de costo del modelo de estimación de esfuerzo: Tipo de Objetivo de Explotación de Información.
P1	Característica del Modelo de Viabilidad perteneciente a la dimensión Plausibilidad que evalúa si los datos disponibles son actuales.
P2	Característica del Modelo de Viabilidad perteneciente a la dimensión Plausibilidad que evalúa si los datos disponibles son representativos.
P3	Característica del Modelo de Viabilidad perteneciente a la dimensión Plausibilidad que evalúa si el problema de negocio es entendible.
P4	Característica del Modelo de Viabilidad perteneciente a la dimensión Plausibilidad que evalúa el nivel de conocimiento del equipo de trabajo.
PEM_E	Esfuerzo estimado por el método empírico de estimación de esfuerzo para Pequeña y Mediana Empresas (en meses/hombre).
PEM_L	Esfuerzo estimado por la fórmula lineal de estimación de esfuerzo para Pequeña y Mediana Empresas (en meses/hombre).
PyME	Pequeña y Mediana Empresa.
QTUA	Factor de costo del modelo de estimación de esfuerzo: Cantidad de Tuplas disponibles en las Tablas Auxiliares.
QTUM	Factor de costo del modelo de estimación de esfuerzo: Cantidad de Tuplas disponibles en la Tabla Principal.
TOOL	Factor de costo del modelo de estimación de esfuerzo: Funcionalidad de las Herramientas Disponibles.
V_d	Valor numérico de la Dimensión *d* para el Modelo de Viabilidad (calculado a partir de I_d).

1. INTRODUCCIÓN

En este capítulo se presenta el contexto de la tesis (sección 1.1), se establecen sus objetivos (sección 1.2), se presentan las publicaciones del tesista vinculadas a las investigaciones realizadas en el desarrollo de la tesis (sección 1.3) y se resume la estructura de la tesis (sección 1.4).

1.1. CONTEXTO DE LA TESIS

Según William Fintan Bohan [2003], hasta el año 1900 la cantidad información que los seres humanos tenían a su disposición se duplicaba cada 1500 años, no obstante a partir del siglo XX la velocidad de generación de nueva información ha aumentado: en 1969 se podía decir que la información se duplicaba cada 50 años; en 1999 cada 5 años y en 2005 cada 39 días. Continuando con dicha progresión, en el año 2020 (es decir en menos de siete años), la información se renovará cada 19 días. Ante esta sobrecarga de información, las organizaciones y la sociedad en general, cada vez necesitan mejores mecanismos para acceder y procesar los datos de manera que estén disponibles, en tiempo y forma, para poder realizar el proceso de toma de decisiones [Kanungo, 2005]. Estos mecanismos deberían poder transformar los datos disponibles mediante su síntesis e interpretación para que sean de utilidad a la persona que debe tomar la decisión.

En este sentido, la Inteligencia de Negocio propone un abordaje interdisciplinario para generar conocimiento que contribuya con la toma de decisiones de gestión y la generación de planes estratégicos en las organizaciones [Thomsen, 2003]. Para ello, cuenta con una sub-disciplina de la Informática denominada Explotación de Información. Dicha disciplina aporta herramientas de análisis y síntesis para extraer el conocimiento no trivial que se encuentra (implícitamente) en los datos disponibles de diferentes fuentes de información [Schiefer *et al.*, 2004].

La Explotación de Información posee semejanzas y diferencias con la Ingeniería de Software. A pesar de tener diferentes objetivos, la Explotación de Información aplica muchos de los principios y buenas prácticas de la Ingeniería de Software para que sus proyectos finalicen satisfactoriamente [Pollo-Cattaneo *et al.*, 2012]. Sin embargo, los métodos, técnicas y herramientas provistas por la Ingeniería de Software no pueden ser reutilizados completamente por estar estos enfocados en características distintas del proyecto [García-Martínez *et al.*, 2011c]; tales como la complejidad de las funciones a implementar, los requerimientos de usabilidad, las interfaces con otros sistemas software, entre otras. Por lo tanto, se detecta la necesidad de desarrollar y validar un conjunto de métodos, técnicas y herramientas que puedan asistir a los practicantes del área de software y proveer la necesaria objetividad, racionalidad, generalización y confiabilidad a los Proyectos de Explotación de Información.

1.2. OBJETIVOS DE LA TESIS

En este contexto, se ha propuesto como objetivo general de este trabajo de tesis definir dos modelos que se pueden aplicar al inicio de un Proyecto de Explotación de Información para dar soporte a su proceso de gestión: un Modelo para la Evaluación de la Viabilidad y un Modelo para la Estimación de Esfuerzo. Para proponer ambos modelos se tienen en cuenta las características de estos proyectos cuando se desarrollan en el ámbito de las Pequeñas y Medianas Empresas.

De esta manera, el primer modelo propuesto en esta tesis se enfoca en la evaluación de las características que se conocen al comienzo de un Proyecto de Explotación de Información; y a partir de esta evaluación, se intenta definir si es posible y adecuado desarrollar el proyecto, así como también predecir si el mismo tendrá éxito. Por otra parte, el segundo modelo tiene como objetivo estimar la cantidad de esfuerzo (en unidades de tiempo/hombre) que serán necesarios para desarrollar el proyecto en forma completa. Como resultado de la aplicación de estos modelos, se espera aportar la información necesaria que le permita a un Ingeniero de Explotación de Información gestionar el proyecto correctamente desde su inicio y así garantizar la finalización exitosa del mismo.

Concretamente, la tesis intenta formular para cada modelo: [a] el conjunto de características que deben ser relevadas y analizadas, y [b] el conjunto de pasos que se deben llevar a cabo para obtener los resultados correspondientes.

1.3. PRODUCCIÓN CIENTÍFICA DERIVADA DE RESULTADOS PARCIALES DE LA TESIS

Durante el desarrollo de esta tesis se han comunicado resultados parciales a través de diversas las siguientes publicaciones:

Capítulos de Libro:

- Pytel, P., Britos, P., García-Martínez, R. 2012. *Initial Activities Oriented to Reduce Failure in Information Mining Projects*. Capítulo 2 en Software Engineering: Methods, Modeling, and Teaching, Volume 2, pp. 11-20. Sello Editorial de la Pontificia Universidad Católica del Perú. ISBN 978-612-4057-84-7.

Artículos en Revistas con Referato:

- Pytel, P., Britos, P., García-Martínez, R. 2013. *Modelos para Asistir la Gestión de Proyectos de Explotación de Información.* Revista Latinoamericana de Ingeniería de Software, 1(1), pp. 8-17. ISSN 2314-2642.
- Pytel, P., Britos, P., García-Martínez, R. 2013. *A Proposal of Effort Estimation Method for Information Mining Projects Oriented to SMEs*. Lecture Notes in Business Information Processing 139, pp. 58-74. ISBN 978-3-642-36610-9.

Comunicaciones a Congresos Internacionales:

- Pytel, P., Britos, P., García-Martínez, R. 2013. *Proposal and Validation of a Feasibility Model for Information Mining Projects*. Proceedings of 25th International Conference on Software Engineering and Knowledge Engineering, pp. 83-88. ISBN 978-1-891706-33-2.

Comunicaciones a Congresos Regionales:

- Pytel, P., Britos, P., García-Martínez, R. 2012. *Comparacion de Métricas de Estimación para Proyectos de Explotación de Información*. Proceedings of Latin American Congress on Requirements Engineering and Software Testing, pp. 29-37. ISBN 978-958-46-0577-1.
- Pytel, P., Amatriain, H., Britos, P., Garcia-Martinez, R. 2012. *Estudio del Modelo para Evaluar la Viabilidad de Proyectos de Explotación de Información*. Proceedings IX Jornadas Iberoamericanas de Ingeniería del Software e Ingeniería del Conocimiento, pp. 63-70. Sello Editorial de la Pontificia Universidad Católica del Perú. ISBN 978-612-4057-85-4.

Comunicaciones a Congresos Nacionales:

- Pytel, P., Tomasello, M., Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. 2011. *Estudio del Modelo Paramétrico DMCoMo de Estimación de Proyectos de Explotación de Información*. Proceedings XVII Congreso Argentino de Ciencias de la Computación, pp. 979-988. ISBN 978-950-34-0756-1.
- Pytel, P., Pollo-Cattaneo, F., Rodríguez, D., Britos, P., García-Martínez, R. 2011. *Identificación de Tareas Críticas en una Metodología de Desarrollo de Proyectos de Explotación*. Proceedings XVII Congreso Argentino de Ciencias de la Computación, pp. 989-998. ISBN 978-950-34-0756-1.

- Pytel, P., Britos, P., García-Martínez, R. 2012. *Viabilidad y Estimación de Proyectos de Explotación de Información*. Proceedings del XIV Workshop de Investigadores en Ciencias de la Computación, pp. 217-221. ISBN 978-50-766-082-5.
- Pytel, P., Britos, P., García-Martínez, R. 2012. *Propuesta de un Modelo para Evaluar la Viabilidad de Proyectos de Explotación de Información*. Proceedings del XVIII Congreso Argentino de Ciencias de la Computación, pp. 1039-1048. ISBN 978-987-1648-34-4.

1.4. VISIÓN GENERAL DE LA TESIS

La tesis se estructura en siete capítulos y cinco anexos que se describen a continuación. Los capítulos incluidos son los siguientes:

En el capítulo "Introducción" se plantea el contexto de la tesis, se establece su objetivo, se presentan las publicaciones del tesista vinculadas a las investigaciones realizadas en el desarrollo de la tesis y se resume la estructura de la tesis.

En el capítulo "Estado de la Cuestión" se presenta la descripción de los principales temas relacionados a este trabajo de tesis. En este sentido, se realiza una introducción sobre los aspectos generales de la Explotación de Información y las características de sus proyectos indicando algunas metodologías que pueden ser aplicadas. En última instancia, se describen dos tipos de modelos aplicados al comienzo de los proyectos de la Ingeniería de Software para asistir en su proceso de gestión: modelos para análisis de viabilidad y modelos para estimación del esfuerzo requerido.

En el capítulo "Descripción del Problema" se identifican los problemas de investigación a ser resueltos en este trabajo de tesis con un sumario de investigación. Estos surgen por las dificultades encontradas normalmente en proyectos de Ingeniería de Software que generan su fracaso, y que también afectan a los proyectos de Explotación de Información.

En el capítulo "Solución" se proponen dos modelos que intentan solucionar los problemas identificados en el capítulo anterior: un Modelo para la Evaluación de la Viabilidad de proyectos de Explotación de Información y un Modelo para la Estimación de Esfuerzo de proyectos pequeños y medianos de Explotación de Información. En primer término, se presenta una introducción sobre los aspectos generales de ambos modelos para luego realizar la propuesta de cada modelo. Cada propuesta incluye un proceso, una lista de características a evaluar y métodos para realizar los cálculos necesarios.

Los modelos propuestos en el capítulo anterior se validan luego en el capítulo de "Validación". Para ello, se aplican ambos modelos en un conjunto de proyectos de Explotación de Información reales que se han recolectados. Posteriormente, los valores obtenidos son comparados con los resultados reales de los proyectos, y analizados por métodos estadísticos.

En el capítulo “Conclusiones” se presentan las aportaciones de esta tesis doctoral y se destacan las futuras líneas de investigación, las cuales se consideran de interés en base al problema abierto que se aborda en este trabajo de tesis.
Finalmente, en el capítulo “Referencias” se listan todas las publicaciones consultadas para el desarrollo de esta tesis.

Por otro lado el contenido de los anexos incluidos en esta tesis es:
En el “Anexo A” se indican los datos de los proyectos de Explotación de Información que fueron utilizados en la regresión aplicada para obtener la fórmula lineal del Modelo de Estimación de Esfuerzo orientado para PyMEs.
Por otro lado, en el “Anexo B” se encuentran detallados los datos de los proyectos de Explotación de Información que fueron utilizados para la validación de los modelos propuestos en esta tesis. A partir de estos datos se aplica el Modelo para la Evaluación de la Viabilidad como se muestra en el “Anexo C”, y los dos métodos del Modelo de Estimación de Esfuerzo en el “Anexo D”. Asimismo, estos datos también son utilizados para demostrar los resultados del modelo de estimación de esfuerzo DMCoMo (“Anexo E”).
Finalmente, en el “Anexo F” se describe la prueba de rangos con signo de Wilcoxon [1945] la cual es aplicada en el capítulo 5 para llevar a cabo la validación de los modelos propuestos.

2. ESTADO DE LA CUESTIÓN

En este capítulo se presenta el estado de los temas relacionados al objetivo de este trabajo de tesis. Para ello, primero se realiza una introducción sobre los aspectos generales de la Explotación de Información (sección 2.1) y las características de sus proyectos (sección 2.2). A posteriori, se describen dos metodologías que pueden ser aplicadas para desarrollar dichos proyectos (sección 2.3). Por último, se describen dos tipos de modelos que pueden ser utilizados al comienzo de un proyecto de Explotación de Información para asistir su gestión (sección 2.4): modelos para análisis de viabilidad (sección 2.4.1) y modelos para estimación del esfuerzo requerido (sección 2.4.2).

2.1. EXPLOTACIÓN DE INFORMACIÓN

La Explotación de Información consiste en la extracción de conocimiento no-trivial que se encuentra distribuido en forma implícita en los datos de las fuentes de información disponibles en una organización [Schiefer *et al.*, 2004]. Dicho conocimiento es previamente desconocido y puede resultar útil para la toma de decisiones dentro de una organización [Thomsen, 2003]. Debido a la gran cantidad de datos que poseen las organizaciones actualmente en sus repositorios, es necesario contar con mecanismos que permitan descubrir relaciones, fluctuaciones y dependencias entre dichos datos [Negash & Gray, 2008]. Si este conjunto de relaciones (o patrones) reflejan la realidad (son válidos) además de aportar algo novedoso y útil para la toma decisiones, entonces es posible indicar que la información toma un grado de valor mayor nivel que el inicial [Kanungo, 2005]. Si se considera la Pirámide de la Información (también conocida como 'DIKW hierarchy') propuesta por [Ackoff, 1989], se puede decir que, en este caso, los datos se transforman (o evolucionan) al nivel de conocimiento como se puede visualizar en la figura 2.1.

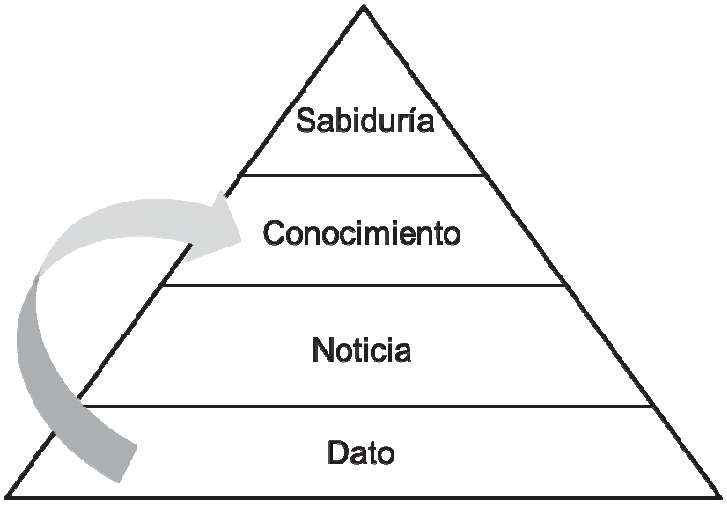


Figura 2.1. Pirámide de la Información.

Para realizar dicha transformación se aplican los denominados Procesos de Explotación de Información [Britos & García-Martínez, 2009]. A partir del problema de negocio que se desea resolver, y las características de los datos disponibles, se selecciona el proceso que mejor se adecue. Cada proceso de Explotación de Información tiene asociado un conjunto de técnicas o algoritmos de Minería de Datos para obtener los resultados necesarios. Muchas de esas técnicas provienen del campo del Aprendizaje Automático [García-Martínez *et al.*, 2003] por lo que los modelos, o patrones de conocimiento, son obtenidos automáticamente. Esto significa que en estos procesos no existe la necesidad de formular hipótesis previas aunque es imprescindible interpretar con cuidado los resultados obtenidos.

Esto marca una diferencia importante entre la Explotación de Información y los sistemas que buscan extraer conocimiento mediante técnicas estadísticas tradicionales (como puede ser los reportes OLAP de un datawarehouse, entre otras implementaciones). Al utilizar técnicas estadísticas tradicionales, primero se debe formular una hipótesis de trabajo de acuerdo al problema de negocio que se quiere resolver, luego se aplican las operaciones necesarias, y con los resultados obtenidos se confirma o refuta la hipótesis propuesta. Por lo tanto, una persona que utiliza estas técnicas debe tener un gran conocimiento sobre el dominio y los datos, ya que la complejidad de los datos almacenados y sus interrelaciones dificulta la verificación del modelo. Al contrario, como se mencionó anteriormente, al aplicar Minería de Datos no es necesario contar con ninguna hipótesis previa ya que estas serán generadas en forma automática por los algoritmos.

Para ilustrar esta diferencia, Hearst [2003] presenta una analogía en el ámbito de la lucha contra el crimen: la diferencia entre el descubrimiento de nuevos conocimientos mediante Minería de Datos, y mediante las técnicas estadísticas tradicionales, es similar a la diferencia existente entre un detective que sigue pistas para encontrar al criminal, contra los analistas que miran las estadísticas del crimen para evaluar las tendencias generales existentes. En forma similar, Pyle [1999] indica que si se imagina que resolver un problema de negocio implica navegar en un vasto océano de datos del cual se intentará pescar (o extraer) el conocimiento necesario para resolverlo, la aplicación del análisis estadístico tradicional es como un niño pescando con una caña a la orilla del océano. En cambio, la Minería de Datos permite obtener el conocimiento necesario con un mediomundo, ya que permite generar diversos patrones de conocimiento en forma directa y automática, para luego ser analizados.

Finalmente, en este punto es necesario realizar una distinción importante entre los términos “Minería de Datos” y “Explotación de Información”. Aunque estos términos suelen utilizarse como sinónimos para referirse al mismo cuerpo de conocimientos, en realidad poseen un alcance diferente [García-Martínez *et al.*, 2011c]. La Minería de Datos se encuentra relacionada con la tecnología (es decir las herramientas y algoritmos) necesaria para transformar los datos en conocimiento mientras

que la Explotación de Información está relacionada con los procesos y las metodologías necesarias para obtener este objetivo. De esta manera, se podría decir que la Minería de Datos está más cercana a la problemática de la programación del software, y la Explotación de Información está más cercana a la Ingeniería de Software.

2.2. PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

Una vez definido el concepto de Explotación de Información y su objetivo, es necesario determinar las características que tienen sus proyectos. En primer término, se describe el proceso general que guía a todo proyecto de esta clase (sección 2.2.1), indicando sus semejanzas y diferencias con los proyectos de la Ingeniería de Software (sección 2.2.2). En segundo lugar, se indican las características más importantes de estos proyectos (sección 2.2.3), las cuales se focalizan en los proyectos implementados por Pequeñas y Medianas Empresas (sección 2.2.4).

2.2.1. Proceso de Descubrimiento de Conocimiento

Para entender un proyecto de Explotación de Información, primero, es recomendable conocer el proceso general que le sirve como marco de referencia. Este proceso fue denominado por primera vez en [Piatetsky-Shaphiro & Frawley, 1991] como Proceso de Descubrimiento de Conocimiento en Bases de Datos (en inglés, Knowledge Discovery in Databases o KDD) por transformar los datos disponibles en patrones de conocimiento [Fayyad *et al.*, 1996]. A los efectos de poder llevar a cabo este proceso, es preciso tener identificado y analizado el problema de inteligencia de negocio que se debe resolver. Una vez conocido el problema de negocio, se realiza una secuencia de actividades en forma iterativa para obtener el conocimiento que le dé solución [Lakshmi & Raghunandhan, 2011]. Estas cuatro actividades primordiales se ilustran en la Figura 2.2.

Figura 2.2. Proceso de Descubrimiento de Conocimiento en Bases de Datos.

A continuación se describen las actividades incluidas en el proceso:

1. *Recolección de los datos*: Consiste en la identificación de las fuentes de información disponibles y la extracción de los datos correspondientes para ser utilizados luego en los pasos posteriores.

2. *Preparación de los datos:* Consiste en modificar los datos obtenidos para que luego se puedan aplicar las técnicas de modelado. Entre las tareas que se pueden realizar se incluyen la integración de los datos (es decir, unificar los datos que pueden venir de diferentes fuentes o repositorios), limpieza de los datos (es decir la corrección o eliminación de datos faltantes, incompletos o con ruido) y formateo de los datos (transformar la estructura o el tipo de los datos según sea necesario). Como se puede ver en la Figura 2.3, esta actividad es normalmente considerada como la que posee mayor esfuerzo dentro de todo el proceso.

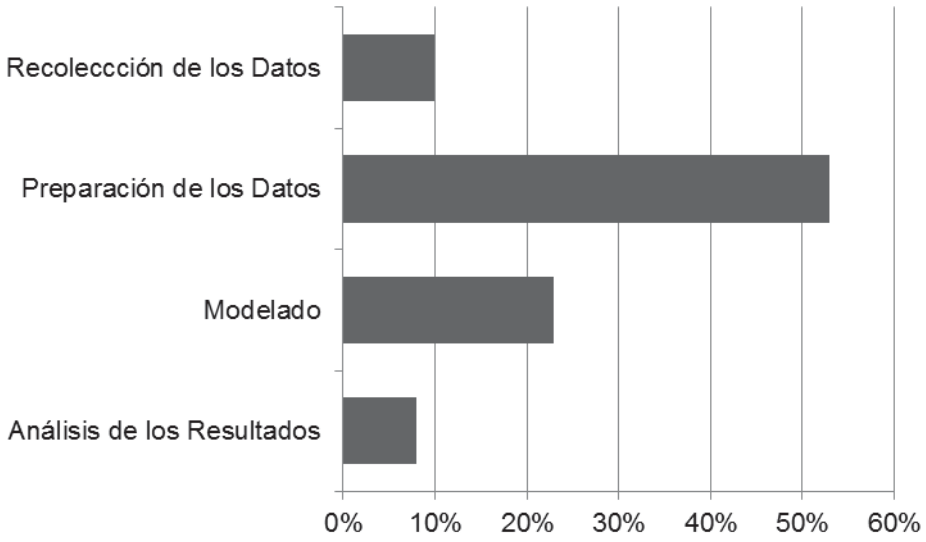


Figura 2.3. Distribución del esfuerzo en el Proceso de Descubrimiento de Conocimiento en Bases de Datos.

3. *Modelado:* Consiste en la aplicación de las técnicas y algoritmos de minería de datos correspondientes para resolver el problema a resolver previamente detectado.

4. *Análisis de los Resultados:* Aunque a veces es subestimada, esta es una de las actividades más importantes, dado que consiste en el estudio de los resultados obtenidos durante el modelado y su interpretación con respecto al dominio de la organización y del problema que se desea resolver. Aquí, los patrones obtenidos son evaluados para corroborar si proporcionan la solución adecuada al problema de negocio, o no. En caso afirmativo, los patrones son presentados y discutidos con los miembros de la organización. No obstante,

cuando no es posible resolver el problema con los resultados obtenidos, entonces será necesario aplicar nuevamente este proceso en forma iterativa.

2.2.2. Relación entre Explotación de Información e Ingeniería de Software

Teniendo en cuenta el Proceso de Descubrimiento de Conocimiento descripto en la sección anterior, se advierte que un proyecto de Explotación de Información posee grandes diferencias con respecto a un proyecto de la Ingeniería de Software. En este sentido, cabe afirmar que la principal diferencia es el objetivo final del proyecto. En un proyecto de Ingeniería de Software, el resultado esperado es un sistema software que brinde solución a las necesidades y deseos de una organización [Pohl, 1997]. Estas necesidades y deseos se expresan en un conjunto de requerimientos (funcionales y no funcionales) que el sistema software deberá satisfacer. En otras palabras, la Ingeniería de Software aplica técnicas, métodos y herramientas para construir un sistema software que cumpla con los requerimientos definidos por la organización dentro de un proyecto que conlleve un tiempo y costo razonable [Sommerville & Sawyer, 1997].

En forma alternativa, en su fase final un proyecto de Explotación de Información obtiene como resultado los patrones de conocimiento a partir de los repositorios de datos disponibles. Para ello, se utilizan herramientas software comerciales o de libre distribución que ya cuentan con un conjunto de operaciones para realizar la preparación y modelados de los datos. Esto significa que un proyecto de Explotación de Información no implica construir un sistema software específico para obtener el conocimiento sino que se pueden usar herramientas ya disponibles. Por consiguiente, en estos proyectos es más importante la gestión de los datos extraídos de los repositorios, el uso de técnicas, métodos y herramientas en el modelado y la correcta interpretación de los resultados para resolver el problema de negocio de la organización. Por otra parte, cabe mencionar que la solución brindada por estos proyectos es de carácter temporal. Una vez que los patrones de conocimiento son utilizados en la toma de decisiones correspondiente, éstos ya no son necesarios. Por el contrario, los sistemas software desarrollados en proyectos de Ingeniería de Software suelen tener una vida operativa mayor, dado que se mantendrán mientras las necesidades que le dieron vida persistan y el contexto en que son utilizados no cambie demasiado.

A pesar de todas estas diferencias, para que un proyecto de Explotación de Información llegue a su finalización de manera satisfactoria, se deben tener en cuentas los principios y las buenas prácticas que se aplican para la Ingeniería de Software. Si se observa la evolución de la Explotación de Información en el tiempo, se puede observar cierto paralelismo con la evolución de la Ingeniería de Software [Mariscal *et al.*, 2007]. En este contexto, se considera un tema abierto la necesidad de organizar un nuevo cuerpo de conocimiento relacionado a una Ingeniería de Explotación de

Información con un foco especial en la implementación y uso en la industria [Pollo-Cattaneo *et al.*, 2012].

2.2.3. Características de los Proyectos de Explotación de Información

Dado que los proyectos de Explotación de Información son un tipo especial de proyecto de Ingeniería de Software se deben identificar sus principales elementos para caracterizarlos. A partir de una investigación documental realizada en [Bolea *et al.*, 2011; Cobos *et al.*, 2010; Davenport, 2009; Fayyad *et al.*, 1996; Fayyad, 2000; García-Martínez *et al.*, 2011c ; Han & Kamber, 2011; Lavravc *et al.*, 2004; Mariscal *et al.*, 2010; Nadali *et al.*, 2011; Nemati & Barkom, 2003; Nie *et al.*, 2009; Pipino *et al.*, 2002; Pyle, 1999; Sim, 2003; Yang *et al.*, 2006] se identifican y definen las principales características de estos proyectos. Dichas características pueden ser clasificadas en función de cuatro aspectos que se describen a continuación:

- La *Organización* donde se realiza el proyecto:
  Al comenzar todo proyecto de Explotación de Información es necesario realizar un relevamiento de la organización donde se está realizando. A tal efecto, es necesario reconocer las características del lenguaje común que poseen las personas de la organización participantes del proyecto [Pollo-Cattaneo *et al.*, 2010]. A su vez, también se debe describir los departamentos que se encuentran relacionados con el proyecto, definiendo su grado de apoyo y compromiso correspondiente.

- El *Problema de Negocio* que se busca resolver:
  Se entiende por problema de negocio a la situación de la organización que dio lugar al comienzo del proyecto de Explotación de Información. Este problema puede ser una situación de crisis (tal como puede ser el caso de una empresa de servicios que está perdiendo clientes) o una nueva oportunidad que se desea aprovechar (como puede ser, por ejemplo, la posibilidad de ofrecer un nuevo servicio o producto). En ambos casos, es necesario que el problema de negocio sea identificado correctamente [Pollo-Cattaneo *et al.*, 2013]. Es decir, se debe entender clara y completamente su alcance, restricciones, expectativas y posibles repercusiones con respecto a la organización. De esta forma, será posible seleccionar los datos disponibles que se deben utilizar y evaluar el proceso de explotación de información a aplicar de acuerdo a los objetivos asociados al problema.

- Los *Datos* disponibles:

  Identificar las fuentes de información disponibles es algo imprescindible para poder realizar un proyecto de Explotación de Información. Si no se cuenta con repositorios con datos actuales, significativos y de buena calidad, no se podrá realizar el proyecto.

  También es recomendable que estos datos se encuentren informatizados; es decir, no sólo en papel impreso sino que también estén en formato digital. En caso contrario, estos datos deberán ser ingresados en un sistema software generando un retraso en el comienzo del proyecto. Actualmente esto no es un impedimento, dado que las organizaciones suelen contar con varios repositorios implementados en diversas tecnologías (como pueden ser sistemas gestores de bases de datos, planillas de cálculos, documentos entre otros). Por lo general, este hecho trae aparejado mayor esfuerzo en las tareas de preparación de datos debido a la necesidad de integrar fuentes tecnológicamente no compatibles.

- Los *Recursos* disponibles para realizar el proyecto:

  Finalmente, se debe determinar los recursos que se cuentan para realizar el proyecto. Estos son los miembros del equipo de trabajo y las herramientas software disponibles.

  En el equipo de trabajo es importante contar con personas que posean conocimientos previos sobre las técnicas, métodos y herramientas a ser aplicadas en el proyecto. En este sentido, será útil contar con personas que posean conocimientos generales sobre Ingeniería de Software, así como conocimientos específicos sobre Explotación de Información y Minería de Datos. También es recomendable que tengan experiencia en proyectos similares, donde se hayan resulto problemas de negocio análogos y/o utilizando datos equivalentes a los disponibles.

  Por otra parte, y para evitar la necesidad de desarrollar las tareas de forma manual o a través de un software generado ad-hoc, es preciso contar al menos una herramienta que incluya funcionalidades relativas a la preparación de los datos y al modelado de los mismos. Afortunadamente, existen disponibles una gran cantidad de herramientas que pueden ser de utilidad, siendo muchas de ellas de código libre y abierto [JMACOE, 2013].

2.2.4. Particularidades de Proyectos de Explotación de Información en PyMEs

Según el informe de las Pequeñas y Medianas Empresas (PyMEs) correspondiente al reporte de espíritu empresarial de la Organización para la Cooperación y el Desarrollo Económico [OECD, 2005], “las PyMEs constituyen la forma dominante de organización empresarial en todos los países

de todo el mundo, representando más del 95% y hasta el 99% de la población de empresas según el país".

Sin embargo, y a pesar de que es bien conocida la importancia de las PyMEs en el contexto internacional, no se tiene conocimiento de que exista un criterio universal para poder identificarlas. Dependiendo de cada país y región se utilizan diferentes criterios cuantitativos y cualitativos para reconocer a una organización como PyME. De esta forma, en Latinoamérica cada país tiene una definición diferente [Álvarez & Durán, 2009]: Argentina identifica como PyME a las empresas autónomas que poseen una facturación menor a u$s 20.000 por año (monto máximo que depende de la actividad realizada); Brasil incluye a todas las compañías con 500 empleados o menos mientras que Colombia considera como PyME a las empresas que poseen hasta 200 empleados y activos menores a los u$s 6.500.

En este sentido, la Organización Internacional para la Estandarización (más conocida como International Organization for Standardization o ISO) ha reconocido la necesidad de especificar diferentes perfiles de ciclos de vida para proyectos de Ingeniería de Software en pequeñas entidades (denominadas en inglés 'Very Small Entities' o VSE), y se encuentra trabajando en el estándar ISO/IEC 29110 [2011]. El término VSE fue definido por el grupo de trabajo 24 de SO/IEC JTC1/SC7 como cualquier "empresa, organización, departamento o proyecto que cuenta con a lo más 25 personas" [Laporte *et al.*, 2008].

A partir de estas definiciones, en este trabajo se contempla que un proyecto de Explotación de Información para PyMEs se encuentra enmarcado como un proyecto realizado en una organización de 250 empleados o menos donde los gerentes de alto nivel (por lo general los propietarios de la empresa) necesitan obtener conocimientos no trivial extraído de las bases de datos disponibles para resolver un problema de negocio específico, sin que existan riesgos especiales en juego. Como normalmente los miembros de la empresa no tienen los conocimientos necesarios, el proyecto es realizado por consultores especializados contratados para llevar adelante el mismo. Asimismo, se puede restringir al equipo del proyecto en un máximo de 15 personas (incluyendo tanto los consultores subcontratados y al personal de la empresa involucrada) para realizar el proyecto en menos de un año.

Las primeras tareas de un proyecto de explotación de información son similares a las de un proyecto de desarrollo de software tradicional, dado que se deben educir las necesidades y deseos de los interesados (stakeholders) de la organización. No obstante, en estos proyectos además se necesita conocer las fuentes de información disponibles en la organización por lo que es preciso relevar los repositorios existentes junto con su estructura. Como estos repositorios suelen no estar correctamente documentados, es preciso entrevistar a los expertos en datos de la organización que pueden ser tanto administradores de las bases de datos como usuarios con gran experiencia en el

manejo de los mismos. Ya que estos expertos son escasos, y con poca disponibilidad, es necesario requerir a su buena disposición (y la de sus supervisores) para que participen en las sesiones de educción y así poder identificar las características de los repositorios a ser utilizados.

Por otro lado, la infraestructura de la Tecnologías de la Información y la Comunicación (TIC) de las PyMEs deber ser analizada. En [Ríos, 2006] se indica que en Latinoamérica más del 70% de las PyMEs cuentan con una infraestructura informática, pero sólo el 37% posee servicios automatizados y/o software propio para realizar sus actividades. En líneas generales hacen uso de aplicaciones comerciales (sobre todo manejadores de planillas de cálculo y de documentos) para registrar su información comercial y operativa. Esto significa que los repositorios a ser utilizados en el proyecto estarán implementados en diferentes formatos y tecnologías. Aunque estos repositorios no suelen ser grandes (normalmente no superan el millón de registros), las tareas de preparación de datos (es decir, limpieza, formateo e integración de los datos) tendrán un esfuerzo considerable.

2.3. METODOLOGÍAS PARA REALIZAR PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

Para el desarrollo de Proyectos de Explotación de información existen varias metodologías que se consideran probadas con un buen nivel de madurez, entre las cuales se destacan CRISP-DM [Chapman *et al.*, 2000], P3TQ [Pyle, 2003] y SEMMA [SAS, 2008]. En este trabajo de tesis sólo se considera y describe la metodología CRISP-DM (sección 2.3.1) por ser considerada la que mejor asiste las tareas de gestión de estos proyectos. Por otra parte, también se incluye una adaptación de la misma, denominada Modelo de Proceso para Proyectos de Explotación de Información (sección 2.3.2), que brinda solución a varios de sus limitaciones y puntos débiles.

2.3.1. Metodología CRISP-DM

A partir de la necesidad de las organizaciones de obtener los patrones de conocimientos encontrados en sus repositorios de datos teniendo en cuenta su utilidad y lineamiento con los objetivos y metas del negocio, se ha propuesto en [Chapman *et al.*, 2000] la metodología CRISP-DM (acrónimo en inglés para 'CRoss Industry Standard Process for Data Mining' o Proceso Estándar para Minería de Datos Independiente de la Industria). Esta metodología se encuentra basada en el Proceso de Descubrimiento de Conocimiento por lo que se puede notar varias semejanzas con la misma; incluyendo además ciertas mejoras entre la que se destaca la caracterización de las características de la organización donde se desarrolla el proyecto con su análisis correspondiente. Al mantener como

foco central los objetivos empresariales del proyecto, se diferencia de otras metodologías que se centran en las características técnicas del desarrollo como es el caso de SEMMA [Cobos *et al.*, 2010].

El principal objetivo de CRISP-DM es permitir desarrollar proyectos mediante un proceso estandarizado (independiente de la industria y la herramienta) y así minimizar los costos que implica un proyecto de este tipo en una organización. La metodología distingue entre cuatro dimensiones diferentes en el contexto del proyecto [Mariscal *et al.*, 2010]:

- El dominio de aplicación es el área específica en la que el proyecto se lleva a cabo.
- El problema de negocio que delimita se los objetivos de los procesos a realizar.
- El aspecto técnico donde se consideran los retos técnicos que deben ser superados para lograr la transformación de los datos.
- La dimensión relacionada con la herramienta y los algoritmos de minería de datos que deben ser aplicados durante el proyecto.

Para cubrir estas dimensiones, la metodología se encuentra estructurada en un proceso con cuatro niveles de abstracción organizados de forma jerárquica en tareas que van desde el nivel más general hasta las más específicas, tal como se puede observar en la Figura 2.4.

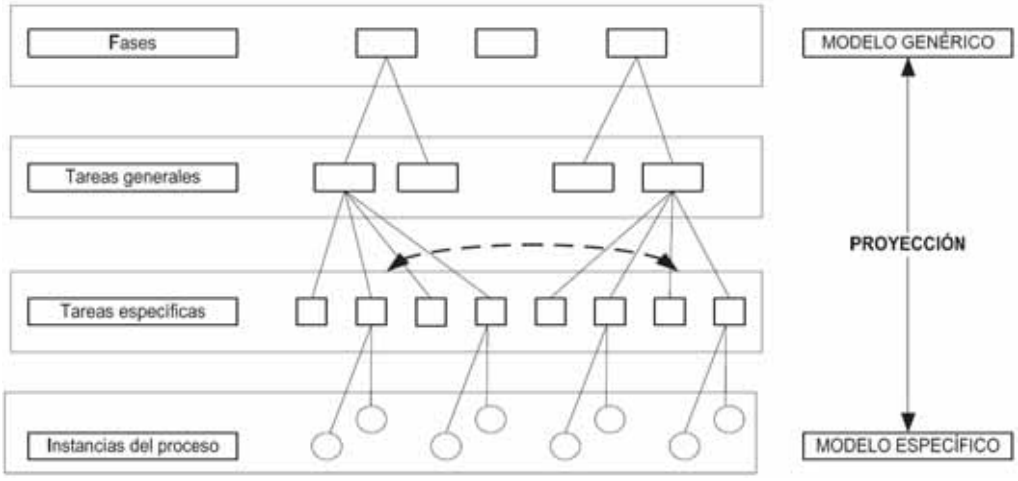


Figura 2.4. Esquema de los cuatro niveles de abstracción de la metodología CRISP-DM.

En su nivel más general, el proceso está organizado en seis fases estando cada fase a su vez estructurada en varias tareas generales de segundo nivel o subfases. Su ciclo de vida se encuentra formado por fases que interactúan entre ellas de forma iterativa (ya que es posible retroceder y volver a una fase anterior si es necesario). Este proceso se puede ver gráficamente en la Figura 2.5 donde el círculo exterior simboliza la naturaleza cíclica del proceso de modelado.

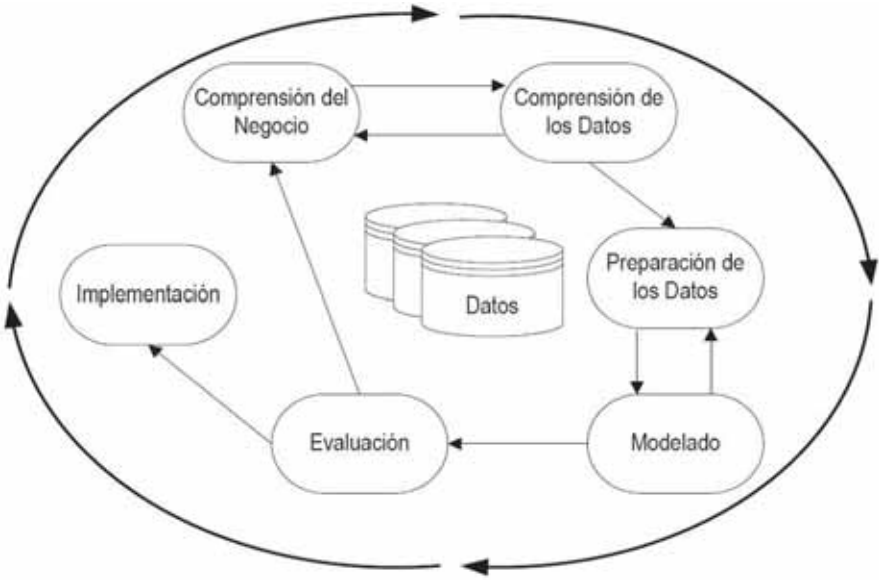


Figura 2.5. Fases del proceso de modelado metodología CRISP-DM.

Cada una de estas fases son descriptas brevemente a continuación:

1. La primera fase es la de '*Comprensión del Negocio*' (o '*Business Understanding*' en inglés) e incluye el entendimiento de los objetivos y requerimientos del proyecto desde una perspectiva de la organización, con el fin de convertirlos en objetivos técnicos y lograr una planificación del resto de las actividades a realizar. Esta fase incluye las tareas que se indican en la Tabla 2.1.

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
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

Tabla 2.1. Tareas de la fase 'Comprensión del Negocio' de la metodología CRISP-DM.

2. La segunda fase, '*Comprensión de los Datos*' (o '*Data Understanding'*), comprende la recolección inicial de los datos necesarios para cumplir los objetivos previamente definidos.

Estos son evaluados, identificando su calidad y estableciendo las relaciones más evidentes. Esto se realiza aplicando las tareas que se indican en la Tabla 2.2.

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
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

Tabla 2.2. Tareas de la fase 'Comprensión de los Datos' de la metodología CRISP-DM.

3. Una vez realizado esta fase, la metodología establece que se proceda a la '*Preparación de los Datos*' ('*Data Preparation*'). Esto significa realizar las tareas de limpieza, formateo e integración sobre los datos recolectados de manera que luego sea posible aplicar las técnicas de modelado. Esta fase se encuentra muy relacionada con la siguiente y ambas interactúan de forma sistemática. Las tareas de esta fase se indican en la Tabla 2.3.

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
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

Tabla 2.3. Tareas de la fase 'Preparación de los Datos' de la metodología CRISP-DM.

4. Durante la fase de '*Modelado*' ('*Modeling'*) se aplican los algoritmos de minería de datos más apropiados, los cuales son determinados por los procesos de explotación de información previamente seleccionados en función de los objetivos y requerimientos del proyecto.
Esto se logra a través de las tareas indicadas en la Tabla 2.4.

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
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

Tabla 2.4. Tareas de la fase 'Modelado' de la metodología CRISP-DM.

5. En la fase de '*Evaluación*' ('*Evaluation'*) se estudia el modelo generado en la fase anterior. A tal efecto, se analiza si el modelo cumple, o no, con los criterios de éxito del problema que fueron identificados en la primera fase. Como resultado de dicho análisis, se puede determinar la necesidad de ejecutar nuevamente alguna de las fases anteriores por haber podido cometer algún error, o pasar a la fase siguiente de '*Implementación*'.
   Las tareas de la fase '*Evaluación*' se encuentran en la Tabla 2.5.

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
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

Tabla 2.5. Tareas de la fase 'Evaluación' de la metodología CRISP-DM.

6. Cuando el modelo generado se considera válido en función de los criterios de éxito, se procede a la '*Implementación*' o '*Deployment*' del proyecto. Ésta incluye la documentación y presentación de los resultados a la organización cliente (Tabla 2.6).

TAREAS GENERALES	TAREAS ESPECÍFICAS ASOCIADAS
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

Tabla 2.6. Tareas de la fase 'Implementación' de la metodología CRISP-DM.

Según la encuesta realizada por [KDnuggets, 2007], CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Explotación de Información como se puede ver en el gráfico de la Figura 2.6. Esta supremacía se mantiene desde el año 2002 y se debe, entre otras razones, a que es de libre distribución (sin costo alguno) y se considera que es la metodología independiente del dominio más efectiva, dado que su alcance incluye todas las complejidades del proyecto a través de tareas fáciles de aplicar. En este sentido, se distingue de la metodología P3TQ que es mucho más compleja.

Sin embargo, aunque se la considera confiable y robusta, entre las principales críticas que se le han realizado, se destaca el hecho que CRISP-DM define *qué hacer* pero no *cómo hacerlo* [Mariscal *et al.*, 2007]. Es decir, indica los entregables a ser preparados por cada tarea pero no formula ningún

tipo de técnica o método específico para realizarla. Esto hecho tiene como consecuencia que muchos equipos de trabajo terminen utilizando adaptaciones y/o metodologías propias.

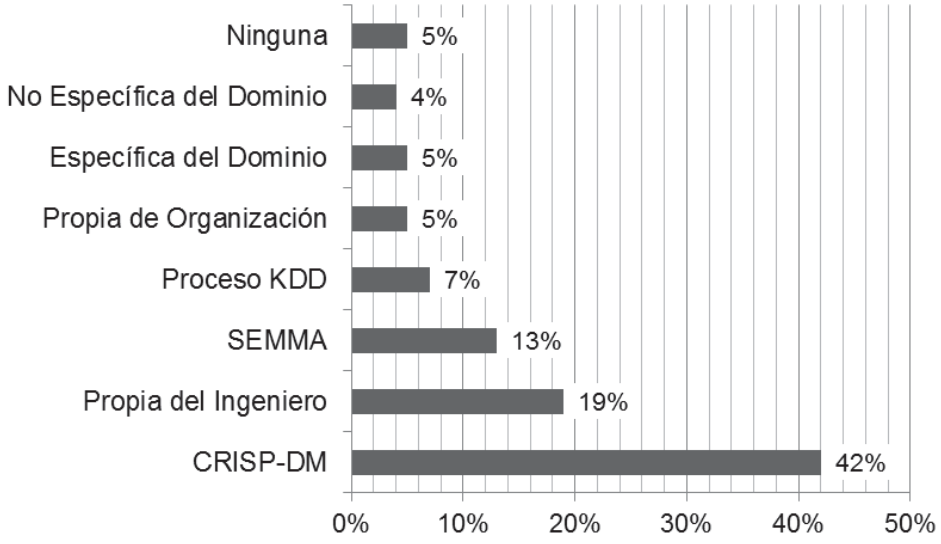


Figura 2.6. Metodologías más utilizadas para proyectos de Explotación de Información.

2.3.2. Modelo de Proceso para Proyectos de Explotación de Información

Debido que la metodología CRISP-DM no incluye actividades relacionadas con la gestión del proyecto (si bien poseen algunos elementos de administración, éstos se encuentran mezclados con los elementos de producción), en [Vanrell *et al.*, 2010; 2012] se propone como solución un Modelo de Procesos para Proyectos de Explotación de Información basado en la combinación de esta metodología con el modelo COMPETISOFT [Oktaba *et al.*, 2007]. COMPETISOFT es la proyección a nivel iberoamericano del modelo de procesos para el desarrollo de software MoProSoft [Oktaba *et al.*, 2003] creado por encargo de la Secretaría de Economía Mexicana para servir de base a la norma Mexicana para la Industria de Desarrollo y Mantenimiento de Software.
Este Modelo de Procesos para Proyectos de Explotación de Información elimina todas las fases no necesarias de CRISP-DM, dejando sólo las que son imprescindibles para realizar la explotación de información y, además, agrega nuevas fases para aspectos específicos de este tipo de proyectos cuando son desarrollados en PyMEs. De esta forma, se incluyen métodos y modelos que permiten controlar la calidad final del producto a desarrollar estableciendo controles sobre cada una de las etapas que intervienen en el proceso productivo. Se entiende 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 organización que lo desarrolla.

Las fases propuestas para el Modelo de Procesos se encuentran divididas en dos procesos. El primero corresponde a las actividades de '*Administración de Proyectos*' que incluye las tareas indicadas en la Tabla 2.7. El otro proceso de '*Desarrollo de Proyectos*' incluye muchas de las tareas de la metodología CRISP-DM (Tabla 2.8).

SUBPROCESO	TAREA	SALIDA
Planificación / Entendimiento del negocio	Entendimiento del negocio	▪ Conocimiento del negocio ▪ Objetivos del negocio ▪ Criterios de éxito
	Definir el proceso específico basado en la descripción del proyecto y el proceso de desarrollo y mantenimiento	▪ Proceso Específico (forma parte del Plan de Desarrollo)
	Definir el protocolo de entrega con el cliente	▪ Plan de Entrega
	Definir ciclos y actividades con base en la descripción del proyecto y en el proceso específico	▪ Proceso Específico (forma parte del Plan de Desarrollo)
	Determinar tiempo estimado para cada actividad	▪ Calendario de actividades (forma parte del Plan de Desarrollo) incorpora el tiempo estimado en el Plan de Proyecto
	Elaborar plan de adquisiciones y capacitación	▪ Plan de Adquisiciones y Capacitación
	Establecer el equipo de trabajo	▪ Equipo de trabajo (forma parte del Plan de Desarrollo)
	Establecer el calendario de actividades	▪ Calendario de actividades (forma parte del Plan de Desarrollo)
	Calcular el costo estimado del proyecto	▪ Costo estimado (forma parte del Plan de Proyecto)
	Evaluación de la situación	▪ Inventario de recursos ▪ Requerimientos, suposiciones y restricciones ▪ Riesgos y contingencias (forma parte del Plan de Proyecto nombrado como Plan de Manejo de Riesgos) ▪ Terminología ▪ Costos y beneficios
	Producir un Plan de Proyecto	▪ Plan de Proyecto, incluye ciclos y actividades, tiempo estimado, plan de adquisiciones y capacitación, equipo de trabajo, costo estimado, calendario, plan de manejo de riesgos y protocolo de entrega
	Producir un Plan de Desarrollo	▪ Plan de Desarrollo (incluye descripción del producto y entregables, proceso específico, equipo de trabajo y calendario) ▪ Lista inicial de técnicas y herramientas
	Formalizar el inicio de un nuevo ciclo del proyecto	
Realización	Acordar las tareas con el equipo de trabajo	
	Acordar la distribución de información	
	Revisar con el responsable la descripción del producto, el equipo de trabajo y el calendario	
	Revisar cumplimiento del plan de adquisiciones y capacitación	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Administrar subcontratos	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Recolectar reportes de actividades y mediciones y sugerencias de mejora y productos de trabajo	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento Reporte de Mediciones y Sugerencias de Mejora
	Registrar costo real del proyecto	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Revisar el registro de rastreo basado en los productos de trabajo recolectados	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Revisar los productos terminados durante el proyecto	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Recibir y analizar las solicitudes de cambio del cliente	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Realizar reuniones con el equipo de trabajo y cliente para reportar avances y tomar acuerdos	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
Evaluación y Control	Evaluar el cumplimiento del plan de proyecto y plan de desarrollo	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Analizar y controlar los riesgos	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Generar el reporte de seguimiento del proyecto	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
Cierre / Entrega	Formalizar la terminación del proyecto o ciclo	▪ Documento de aceptación
	Llevar a cabo el cierre del contrato con subcontratistas	
	Generar el reporte de mediciones y sugerencias de mejora	▪ Reporte de mediciones y sugerencia de mejoras ▪ Lecciones Aprendidas
	Planear la entrega	▪ Plan de entrega (forma parte del Plan de Proyecto nombrado como protocolo de entrega)

Tabla 2.7. Proceso de Administración de Proyectos del Modelo de Procesos.

SUBPROCESO	TAREA	SALIDA
Entendimiento del negocio	Determinar las metas del Data Mining	▪ Metas del Data Mining
		▪ Criterios de éxito del Data Mining
Entendimiento de los datos	Reunir los datos iniciales	▪ Reporte de datos iniciales
	Describir los datos	▪ Reporte de descripción de datos
	Explorar los datos	▪ Reporte de exploración de 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 Data Mining 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 2.8. Proceso de Desarrollo de Proyectos del Modelo de Procesos.

2.4. MODELOS PARA ASISTIR EL INICIO DE UN PROYECTO DE EXPLOTACIÓN DE INFORMACIÓN

En esta sección se presenta las características de los modelos existentes que permiten asistir la gestión de un proyecto de Explotación de Información desde su comienzo: modelos para análisis de viabilidad (sección 2.4.1) y modelos para estimación del esfuerzo requerido (sección 2.4.2).

2.4.1. Análisis de Viabilidad en Proyectos de Explotación de Información

Al comienzo de todo proyecto software, la organización debe decidir si es conveniente realizarlo o no. Para poder tomar esa decisión, la cual es compleja y depende de gran cantidad de factores, se necesita conocer el impacto que ese software va a causar en la organización y los riesgos con los que ella corre debido a su construcción [Sommerville, 2005]. En este sentido, es necesario estudiar las características del proyecto a través de una evaluación de la viabilidad técnica y económica del proyecto (también conocida como estudio de factibilidad). Como resultado de esta evaluación se puede determinar si se cumplen las condiciones para garantizar la finalización del desarrollo del sistema software de manera satisfactoria.

Para llevar a cabo la evaluación de la viabilidad técnica de un proyecto software tradicional, en primer lugar se debe recolectar información tanto sobre la organización como del sistema a desarrollar. Entre los principales factores a establecer se destacan la actitud de los directivos y de

los usuarios hacia el proyecto, el nivel de experiencia de los desarrolladores, el alcance del problema a resolver, y la integración con otros sistemas software, entre otros. Toda esta información recolectada luego es procesada mediante el método correspondiente generando así un informe que indica el grado de viabilidad técnica del proyecto.

En los proyectos de construcción de sistemas basados en conocimiento sucede algo similar; con la diferencia que en la Ingeniería del Conocimiento (INCO) se debe evaluar la viabilidad del proyecto considerando varias dimensiones [Gómez *et al.*, 1997]. Dado que las especificaciones iniciales de estos sistemas suelen ser inciertas, incompletas y contradictorias, es necesario desarrollar distintos prototipos para definir coherentemente las funcionalidades, el rendimiento y las interfaces del sistema [García-Martínez *et al.*, 2003]. Esto produce que los proyectos de la INCO sean más largos y costosos que los de software tradicional [García-Martínez & Britos, 2004]. En este sentido, la metodología IDEAL incluye un test de viabilidad de tipo métrico que busca evaluar si un proyecto es posible, adecuado, justificado y va a tener éxito mediante la manipulación de valores lingüísticos, a través de su representación mediante intervalos difusos que son procesados y combinados, obteniendo de esta manera el valor de viabilidad final del proyecto.

Los Proyectos de Explotación de Información cuentan con una necesidad similar de evaluar la viabilidad del proyecto antes de comenzar el mismo. De esta forma, es posible detectar los problemas asociados al inicio del proyecto y así se podrían reducir los riesgos durante el desarrollo del proyecto. No obstante, dado que las características de los proyectos de explotación de información son diferentes a los proyectos de software tradicional y a los proyectos de la INCO, no es posible reutilizar los modelos propuestos para estos tipos de proyectos, por lo que es necesario contar con modelos específicos.

En virtud del análisis realizado en el marco de este trabajo de tesis se han encontrado diversos trabajos [Bolea *et al.*, 2011; Davenport, 2009; Fayyad, 2000; Lavravc *et al.*, 2004; Nemati & Barkom, 2003; Pipino *et al.*, 2002; Sim, 2003] cuyo objetivo es la identificación de los criterios de éxitos de estos proyectos. Por otro lado, [Nadali *et al.*, 2011] propone un modelo que utilizando un sistema experto difuso [Jang, 1997] permite medir el nivel de éxito de proyectos a partir de la calidad empleada en cada una de sus fases de CRISP-DM. Sin embargo, este último estudio, sólo puede ser aplicado una vez que el proyecto ya está finalizado por necesitar conocer el nivel de calidad empleado en cada fase.

En cambio, [Nie *et al.*, 2009] aplica un análisis Bayesiano [Berger, 1985] para determinar si la empresa se encuentra calificada para comenzar un proyecto de Explotación de Información. Aunque se valoran las características de la empresa y los datos disponibles para decidir si se puede aplicar este tipo de proyecto o no, este estudio no considera las características del problema de negocio y

gestión del proyecto. Además, este análisis deja de lado la clasificación de la viabilidad en diferentes dimensiones por lo que considera el aspecto de la viabilidad como un todo.
Por consiguiente, es posible afirmar que los trabajos analizados no brindan un mecanismo completo para evaluar la viabilidad de un proyecto de Explotación de Información al inicio del mismo.

2.4.2. Estimación de Esfuerzo en Proyectos de Explotación de Información

Una vez que el proyecto es considerado como viable, es preciso realizar las actividades de planificación del proyecto. Estas actividades necesitan una apreciación del posible esfuerzo requerido del proyecto por lo cual es necesario realizar una estimación del trabajo a ejecutar, de los recursos necesarios y del tiempo que transcurrirá desde el comienzo hasta el final del mismo [Pressman, 2005].
En el ámbito de la Ingeniería de Software esto ha llevado a la construcción de métodos de estimación que logren resultados predictivos sobre los recursos a emplear y que se ajusten de la mejor manera posible a la realidad obtenible es un problema abierto en el campo de los sistemas de información [Rodríguez *et al.*, 2010]. El proceso de estimación de un proyecto software está formado por un conjunto de técnicas y procedimientos utilizados en la organización para poder llegar a una predicción fiable. Éste es un proceso continuo, que debe ser usado y consultado a lo largo de todo el ciclo de vida del proyecto [Agarwal & Kumar, 2001]. Se puede decir que toda técnica de estimación de un proyecto software pertenece a una de estas tres categorías [Bielak, 2000]:

a) Modelos Analíticos que a través de la aplicación de métodos de regresión en datos históricos describen las relaciones matemáticas entre las variables del proyecto. Entre estos modelos se destacan el método “COnstructive COst MOdel” o COCOMO [Boehm, 1984] y el método COCOMO II [Boehm *et al.*, 2000].

b) Técnicas basadas en Teorías que consideran las bases teóricas del proceso de desarrollo de software. Entre estos modelos se destaca el método “Software LIfecycle Management” o SLIM [Putman, 1978].

c) Técnicas Empíricas que incluyen la observación y entendimiento de esfuerzos históricos de proyectos concluidos para predecir esfuerzos futuros. Se destaca la técnica de Puntos de Función [Albrecht & Gaffney, 1983; Fairley, 1992] entre estos modelos.

Los Proyectos de Explotación de Información no escapan a la necesidad de estimar el esfuerzo necesario para la construcción del sistema y la planificación de las actividades necesarias. En la metodología CRISP-DM requiere de un proceso de planificación que se encuentra definido dentro de la fase de 'Comprensión del Negocio', pero no propone ningún mecanismo, técnica ni herramienta para realizar dicha estimación. Dada las diferencias que existen entre un proyecto de construcción de software tradicional y de explotación de información, los métodos de estimación disponibles para proyectos de software tradicional no son aplicables ya que los parámetros a ser utilizados son de naturalezas diferentes [Marbán *et al*., 2008]. Por ejemplo, las herramientas de estimación de software tradicional utiliza como parámetros la cantidad de líneas de código, la experiencia del equipo de trabajo, características de la plataforma de desarrollo, entre otras. Estos parámetros tiene sentido en un proyecto de Explotación de Información dado que estos proyectos no siempre necesitan implementar un software para obtener la solución. Para estos proyectos existen otros parámetros que se deben analizar para obtener una estimación confiable, como por ejemplo, cantidad de fuentes de información, nivel de integración de los datos, el tipo de problema a ser resuelto, entre las más representativas. Por lo tanto, y al no poder reutilizar los métodos existentes en otras ingenierías, es necesario contar con métodos específicos de estimación de esfuerzo para proyectos de Explotación de Información.

Luego de una búsqueda documental se ha encontrado dos modelos que podrían aplicar en proyectos de este tipo, uno de naturaleza empírica y el otro analítico:

- Para obtener un *modelo empírico de estimación* en [Rodríguez *et al.*, 2010] se han utilizado registros de proyectos realizados por alumnos universitarios, con sus tamaños estimados en forma temprana y el esfuerzo de desarrollo (sin incluir la puesta en marcha ni su mantenimiento) medido en tiempo/hombre. Con ello se busca obtener la distribución porcentual de la carga de trabajo en proyectos de Explotación de Información realizados por pequeños y medianos emprendimientos.

- En cuanto al modelo analítico [Marbán *et al*., 2008], que se denomina '*Modelo Matemático Paramétrico de Estimación para Proyectos de Data Mining*' o DMCoMo por su nombre en inglés (*Data Mining COst MOdel*). Este es un modelo de estimación de esfuerzo paramétrico de la familia de COCOMO II el cual permite estimar los meses/hombre que serán necesarios para el desarrollo a partir de una serie de variables (denominadas factores de costo) vinculadas a las características más importantes de los proyectos de explotación de información. Estos factores de costo, junto con la categoría a la que pertenecen, son indicados en la Tabla 2.9.

CATEGORÍA	DESCRIPCIÓN	FACTORES DE COSTO
Relacionados a los Datos	Agrupa los factores de costo que tienen que ver con la cantidad, la calidad de los datos a tratar en el proyecto de explotación de información.	– Cantidad de Tablas (NTAB) – Cantidad de Tuplas de las Tablas (NTUP) – Cantidad de Atributos de las Tablas (NATR) – Grado de Dispersión de los Datos (DISP) – Porcentaje de valores NULL (PNUL) – Grado de Documentación de las Fuentes de Información (DMOD) – Grado de Integración de Datos Externos (DEXT)
Relacionados a los Modelos	Incluye todos aquellos factores de costo 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 (NMOD) – Tipo de Modelos a ser Creados (TMOD) – Cantidad de Tuplas de los Modelos (MTUP) – Cantidad y Tipo de Atributos por cada Modelo (MATR) – Cantidad de Técnicas Disponibles para cada Modelo (MTEC)
Relacionados al Desarrollo de la Plataforma	Agrupa los factores de costo que tienen que ver con las características de los almacenes de datos y su localización.	– Cantidad y Tipo de Fuentes de Información Disponibles (NFUN) – Distancia y Medio de Comunicación entre Servidores de Datos (SCOM)
Relacionados a las Técnicas y las Herramientas	Agrupa las características de las técnicas y herramientas de explotación de información que se van a utilizar en el proyecto.	– Herramientas Disponibles para ser Usadas (TOOL) – Grado de Compatibilidad de las Herramientas con Otros Software (COMP) – Nivel de Formación de los Usuarios en las Herramientas (NFOR)
Relacionados al Proyecto	Agrupa aquellas características relativas a los departamentos y a las localizaciones para las que se desarrolla el proyecto de explotación de información.	– Cantidad de Departamentos Involucrados en el Proyecto (NDEP) – Grado de Documentación que es necesario generar (DOCU) – Cantidad de Sitios donde se realizará el Desarrollo y su Grado de Comunicación (SITE)
Relacionados al Equipo de Trabajo	Incluye aquellos factores relacionados con el equipo de trabajo que participa en el proyecto (dirección, implementadores, expertos, etc.).	– Grado de Familiaridad con el Tipo de Problema (MFAM) – Grado de Conocimiento de los Datos (KDAT) – Actitud de los Directivos (ADIR)

Tabla 2.9. Factores de Costo considerados por el modelo DMCoMo.

Una vez que los valores de los factores de costo son conocidos, se ingresan en las ecuaciones matemáticas suministradas por el método. DMCoMo dispone de dos fórmulas, una que utiliza 23 factores de costo (MM23) y otra de 8 factores de costo como variables (MM8). Estas fórmulas se detallan en la Tabla 2.10.

MM23 = 78,752 + 2,802 x NTAB + 1,953 x NTUP + 2,115 x NATR + 6,426 x DISP + 0,345 x PNUL + (- 2,656) x DMOD + 2,586 x DEXT + (-0,456) x NMOD + 6,032 x TMOD + 4,312 x MTUP + 4,966 x MATR + (-2,591) x MTEC + 3,943 x NFUN + 0,896 x SCOM + (-4,615) x TOOL + (-1,831) x COMP + (-4,689) x NFOR + 2,931 x NDEP + (-0,892) x DOCU + 2,135 x SITE + (-0,214) x KDAT + (-3,756) x ADIR + (-4,543) x MFAM	MM8 = 70,897 + 2,368 x NTAB + 2,885 x NATR + 4,792 x DISP + 2,713 x DEXT + 7,257 x TMOD + 4,615 x MATR+ (-3,842) x NFOR + (-3,275) x MFAM

Tabla 2.10. Fórmulas utilizadas por el modelo DMCoMo.

Sin embargo, tal como lo indican sus autores, el método se considera confiable sólo para estimar el esfuerzo de proyectos que se encuentren en el rango de esfuerzo de 90 a 185 meses/hombre (es decir aproximadamente 7,50 a 15,40 años/hombre). Según el estudio estadístico realizado en [Pytel *et al.*, 2011] se ha detectado que este modelo de estimación sólo es confiable para proyectos grandes (es decir proyectos con un esfuerzo estimado mayor a 7 años/hombre). Esto significa que este modelo de estimación no es confiable para proyectos pequeños que son los que usualmente requieren las PyMEs.

3. DESCRIPCIÓN DEL PROBLEMA

En este capítulo se presenta una breve descripción de las dificultades que generan el fracaso de proyectos de Ingeniería de Software (sección 3.1) las cuales también afectan a los proyectos de Explotación de Información (sección 3.2). A partir de dichos problemas se identifica el problema de investigación a ser resuelto en este trabajo de tesis (sección 3.3) y se concluye con un sumario de investigación (sección 3.4).

3.1. ANÁLISIS DEL FRACASO DE PROYECTOS DE INGENIERÍA DE SOFTWARE

El proceso de desarrollo definido por la Ingeniería de Software posee como objetivo principal proveer un sistema software que incluya todas las funcionalidades y características requeridas por el cliente, en el tiempo y costo previamente acordado. Sin embargo, la mayoría de los proyectos de software tradicional pueden ser considerados, al menos, como fracasos parciales debido a que pocos proyectos cumplen con sus presupuestos de costo, planificación, criterios de calidad o especificaciones de requerimientos [May, 1998]. El grupo Standish emite anualmente un 'Reporte del Caos' [Standish Group, 1995] en el que se ilustra el estado actual del desarrollo de software mediante la presentación de estadísticas relacionadas a los proyectos realizados en todo el mundo y las cuales son analizadas y discutidas por dicho grupo. En virtud de lo expuesto los proyectos son clasificados en tres tipos:

- *Proyectos Exitosos* donde se incluye aquellos que son completados en tiempo, con el presupuesto planificado e incluyendo todas las funcionalidades requeridas.
- *Proyectos Cuestionados* los cuales se logran completar al generar un software operacional pero sobrepasando los tiempos y presupuesto estimados y/o no incluyendo todas las funcionalidades requeridas.
- *Proyectos Fallados* que se cancelan antes de ser completado o nunca es implementado.

En la Figura 3.1 se muestra la distribución de estos tipos de proyectos desde el año 1994 hasta el 2010 de acuerdo a las estadísticas recolectadas en [Standish Group, 2010]. Como se puede observar en estos 16 años la proporción de proyectos exitosos ha aumentado en más del doble. Esto es logrado gracias a las buenas prácticas y metodologías, junto con sus métodos, técnicas y

herramientas asociadas, que la comunidad de investigadores en la Ingeniería de Software ha estado proponiendo y utilizando a lo largo de este período.

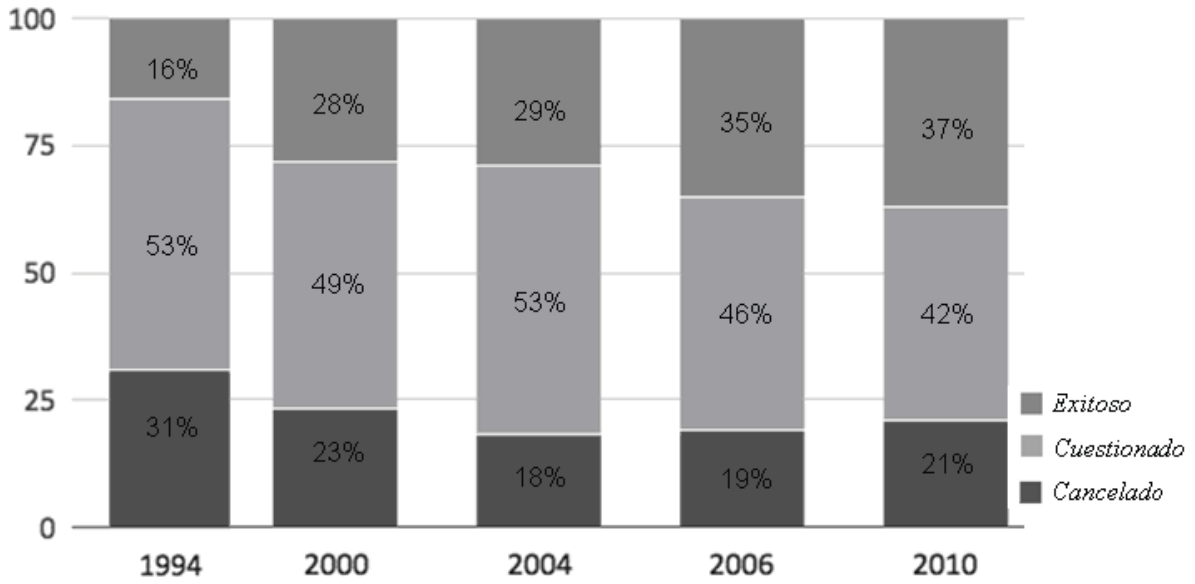


Fig. 3.1. Resultados del estado final de proyectos de software tradicional según el grupo Standish.

No obstante, como también se puede notar en la Figura 3.1, este trabajo no está terminado: en el año 2010, todavía la mayoría de los proyectos (aproximadamente un 63%) han finalizado con problemas o han sido cancelados antes de su finalización. En promedio, la mitad de estos proyectos, han superado el presupuesto acordado en un 190% y han tenido un retrasado del 220% con respecto a su planificación original. En una encuesta realizada por el grupo Standish, se identifican las siguientes causas para estos problemas:

- Requerimientos incompletos (13,1%)
- Pobre inclusión de los usuarios (12,4%)
- Fallas en planificación y estrategia (10,6%)
- Expectativas no realistas (9,9%)
- Falta de Soporte Gerencial (9,3%)
- Requerimientos y especificaciones cambiantes (8,7%)
- Falta de Planificación y recursos insuficientes (8,1%)
- Requerimientos que dejan de ser necesarios (7,5%)
- Pobre Manejo de IT (6,2%)
- Desconocimiento de la Tecnología (4,3%)

Como se puede ver, para el grupo Standish los tres principales causas están asociadas al manejo de los requerimientos (aproximadamente el 41,7%), la planificación de tiempo y recursos necesarios para desarrollar el proyecto en forma completa (el 18,7% aproximadamente), y el contar con objetivos no realistas o expectativas inalcanzables (en un 9,9% de los casos). Esto se confirma en [Charette, 2005] donde se identifican las siguientes causas que originan el fracaso de los proyectos:

- Metas del proyecto no realistas.
- Estimaciones inexactas de los recursos necesarios.
- Requisitos del sistema mal definidos.
- Deficiente información de estado del proyecto.
- Riesgos no identificados ni controlados.
- Falta de comunicación entre los clientes, desarrolladores y usuarios.
- Uso de tecnología inmadura.
- Incapacidad para manejar la complejidad del proyecto.
- Prácticas de desarrollo descuidadas.
- Mala gestión del proyecto.
- Política de los stakeholders.
- Presiones comerciales.

De nuevo aparecen los problemas en los requerimientos, pero también se detectan problemas relacionados con la falta de información necesaria para la gestión correcta del proyecto. En este sentido, se destaca nuevamente la falta de una correcta estimación de esfuerzo para poder determinar los recursos necesarios que necesita el proyecto y la presencia de metas o expectativas no alcanzables. Ambos problemas pueden verse combinados en las tres primeras leyes de Golub sobre proyectos informáticos [Bloch, 2003]:

i. Los proyectos con objetivos difusos, van bien sólo para evitar el compromiso de tener que estimar los costos.

ii. Un proyecto planificado sin precisión tarda tres veces más en acabarse de lo que se espera, un proyecto planificado cuidadosamente tarda el doble de lo previsto.

iii. El esfuerzo requerido para corregir el curso de un proyecto se incrementa geométricamente en función del tiempo transcurrido.

3.2. ANÁLISIS DEL FRACASO DE PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

Teniendo en cuenta que los proyectos de Explotación de Información constituyen un tipo especial de proyectos en el campo de la Ingeniería de Software, los problemas que posee son similares [García-Martínez *et al*., 2011c]. Estudios realizados sobre proyectos de Explotación de Información han detectado que la mayoría finalizan con fracasos [Edelstein & Edelstein, 1997; Strand, 2000]. En el año 2000, se había determinado que el 85% de los proyectos no alcanzan sus metas [Fayyad, 2000], mientras que en el 2005 el porcentaje de fracaso ha bajado a aproximadamente el 60% [Gondar, 2005] hasta alcanzar el 50% en el año 2009 [Marbán *et al.*, 2009]. Esto parece indicar que la comunidad ha estado trabajando en el camino correcto. Mediante la propuesta, desarrollo y validación de métodos, técnicas y herramientas, se intenta lograr una mejora en el resultado final del proyecto. Los métodos con abordaje ingenieril permiten dotar al proceso de desarrollo de los siguientes aspectos: objetividad, sistematicidad, racionalidad, generalidad y fiabilidad, contribuyendo al avance del conocimiento científico mediante la definición de técnicas consistentes [Pollo-Cattaneo *et al.*, 2012].

Dentro de estas propuestas, se puede destacar el Modelo de Proceso para la Gestión de Requerimientos definido en [Pollo-Cattaneo *et al.*, 2013a; 2013b]. Este modelo de proceso guía al Ingeniero de Explotación de Información en las actividades de elicitación y documentación de los requerimientos del proyecto. Dentro de estos requerimientos se establecen los objetivos del proyecto, el vocabulario del negocio, las fuentes de información disponibles y un conjunto de procesos de Explotación de Información [Britos & García-Martínez, 2009] que buscan solucionar los problemas de negocio que dieron origen al proyecto. Esta información luego es utilizada tanto por las actividades de Gestión de Proyecto como el Proceso de Descubrimiento de Conocimiento en Bases de Datos [Piatetsky-Shaphiro & Frawley, 1991]. Esta relación entre el modelo de proceso, las actividades de gestión y el proceso de descubrimiento de conocimiento, para proyectos de Explotación de Información, se puede observar en la Figura 3.2.

Sin embargo, todavía existen cuestiones de gestión que deben ser mejoradas en el marco de los proyectos de Explotación de Información. Si bien existen modelos y metodologías que acompañan su desarrollo, las cuales se consideran probadas con un buen nivel de madurez, éstas dejan de lado aspectos a nivel operativo de los proyectos y de empresa [Vanrell *et al.*, 2010]. En este sentido, se *destaca la ausencia de procesos y herramientas que permitan soportar muchas de las* ***actividades de gestión sobre todo al comienzo del proyecto***. Estas actividades son de gran importancia para reducir la probabilidad de fracasos en el desarrollo de estos proyectos.

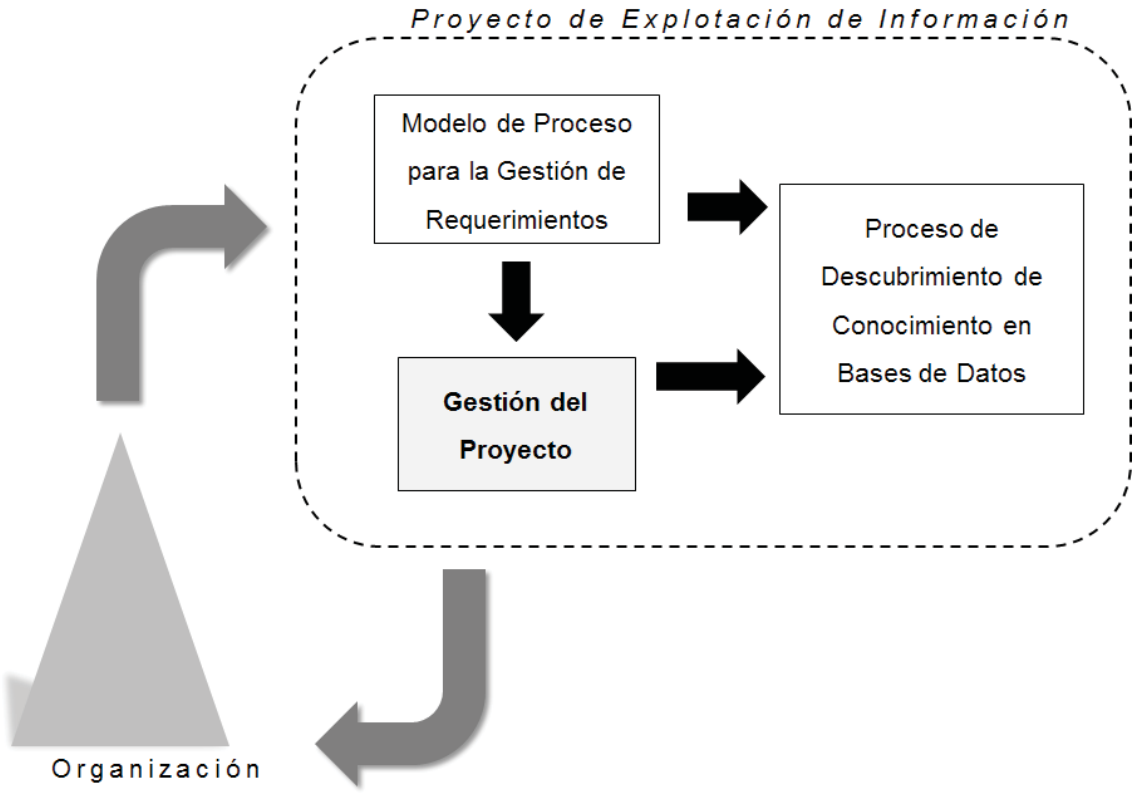


Fig. 3.2. Relación entre la gestión de requerimientos, gestión del proyecto y el proceso de descubrimiento de conocimiento en Bases de Datos en proyectos de Explotación de Información.

3.3. IDENTIFICACIÓN DEL PROBLEMA DE INVESTIGACIÓN

A los efectos de identificar el problema que se pretende resolver en el presente trabajo de tesis, se considera la carencia de modelos que puedan ser aplicados en las fases iniciales de las metodologías existentes para proyectos de Explotación de Información. Al considerar las principales causas de fracasos para proyectos de Ingeniería de Software y de Explotación de Información, se pone de manifiesto la necesidad de proveer mecanismos que permitan al Ingeniero de Explotación de Información responder ciertos interrogantes asociados a la gestión inicial del proyecto. Por una parte, es necesario establecer si es posible realizar el proyecto, si la solución es la correcta y si es posible cumplir con todas las metas y expectativas del proyecto; es decir, si el mismo puede concluir de manera exitosa. Por otra parte, es preciso conocer la cantidad de personas que van a participar en el equipo de trabajo durante el desarrollo del proyecto, y por cuanto tiempo. Estos interrogantes se representan gráficamente en la Figura 3.3.

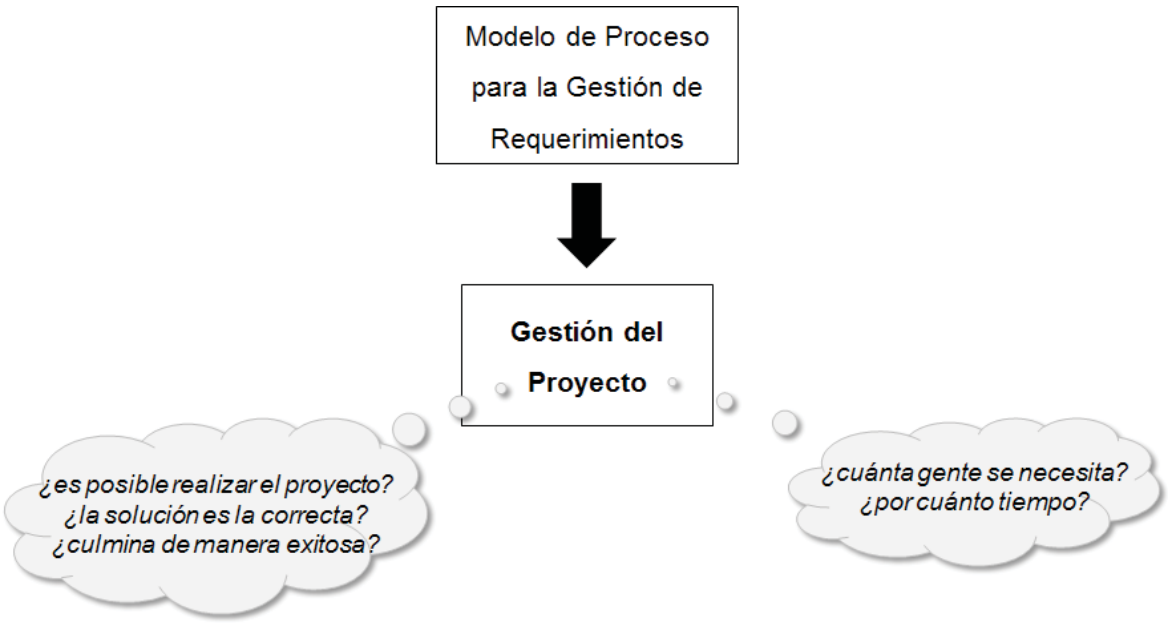


Fig. 3.3. Interrogantes asociados a la gestión inicial del proyecto.

En virtud de lo expuesto, se hace necesario evaluar las características principales de la organización y del proyecto, teniendo en cuenta sólo aquellas que son conocidas al comienzo del mismo. Dado que los proyectos de Explotación de Información presentan diferencias importantes con los proyectos de Software Tradicional y con los proyectos de la Ingeniería del Conocimiento, no es posible reutilizar los modelos disponibles para este tipo de proyectos. Por consiguiente, es preciso disponer de modelos específicos que tengan en cuenta las particularidades de los proyectos de Explotación de Información. En tal sentido, estos modelos deben incluir la evaluación de las metas del proyecto, las expectativas y la relación de los interesados (stakeholders, en inglés), particularidades de la organización y de los datos disponibles, así como el tipo de tecnología a ser aplicada; entre otros aspectos.

Asimismo cabe destacarse que, tanto en Argentina como en Latinoamérica, existe gran cantidad de Pequeñas y Medianas Empresas (PyMEs) las cuales poseen características especiales que las distinguen de empresas más grandes. En este sentido, se puede señalar la escasa cantidad de trabajos existentes en la literatura científica sobre modelos que se pueden aplicar en proyectos de Explotación de Información realizados en este tipo de empresas. En consecuencia, se estima que es de interés disponer de métodos que habiliten al sector PyMEs a implementar servicios de la Inteligencia de Negocios que contribuyan a la toma de decisiones en los niveles de gestión de la industria y el comercio regional [García-Martínez *et al.*, 2011b].

En este contexto, el presente trabajo de tesis tiene como objetivo proponer, estudiar y validar dos modelos a ser utilizados por las PyMEs al inicio de un proyecto de Explotación de Información,

contribuyendo así a disminuir las distintas dificultades que se pueden presentar. El primer modelo propuesto permite realizar la ***Evaluación de la Viabilidad*** del proyecto, a los efectos de determinar de manera temprana los puntos débiles asociados a aquellas características del proyecto que pueden presentar dificultades durante su desarrollo; mientras que el segundo modelo, tiene como objetivo realizar la ***Estimación del Esfuerzo*** (medido en cantidad de personas por tiempo) que se requiere para llevar a cabo el proyecto de manera satisfactoria y así poder planificar las actividades que se deben desarrollar. Ambos modelos son descriptos en el capítulo 4 correspondiente a la Solución del problema.

3.4. SUMARIO DE INVESTIGACIÓN

De lo expuesto precedentemente surgen las siguientes preguntas de investigación:

Pregunta 1: ¿Es posible determinar si la realización de un proyecto de Explotación de Información, que se desarrolla en el ámbito de las PyMEs, es viable a partir de las características que se identifican al inicio?

En caso afirmativo, surgen dos preguntas adicionales:

Pregunta 1a: ¿Qué características del proyecto se deben estudiar para determinar su viabilidad?

Pregunta 1b: ¿Cuáles son los pasos que se debe realizar para llevar a cabo dicho estudio y así detectar los puntos débiles asociados al desarrollo del proyecto?

Pregunta 2: ¿Es posible predecir de manera temprana el esfuerzo para desarrollar un proyecto de Explotación de Información en forma completa, teniendo en cuenta las características requeridas por las PyMEs?

En caso afirmativo, surgen dos preguntas adicionales:

Pregunta 2a ¿Qué características se deben considerar a los efectos de estimar la cantidad de personas y el tiempo necesario que insume el proyecto?

Pregunta 2b: ¿Cuáles son los métodos que se pueden emplear para obtener la estimación de dicho esfuerzo?

En los próximos capítulos de este trabajo se proponen soluciones a los interrogantes planteados y su correspondiente validación.

4. SOLUCIÓN

En este capítulo se presentan los modelos que buscan solucionar los problemas identificados en el capítulo anterior. En primer lugar, se realiza una introducción sobre los aspectos generales de ambos modelos (sección 4.1). En segundo término, se describe el Modelo para la Evaluación de la Viabilidad de Proyectos de Explotación de Información (sección 4.2), del cual se abordan las cuestiones generales de mayor relevancia (sección 4.2.1), la propuesta correspondiente (sección 4.2.2) y su análisis preliminar (sección 4.2.3). Finalmente, se describe el Modelo para la Estimación de Esfuerzo de Proyectos de Explotación de Información (sección 4.3), del cual también se abordan las cuestiones generales de mayor relevancia (sección 4.3.1), la propuesta correspondiente (sección 4.3.2) y su análisis preliminar (sección 4.3.3).

4.1. INTRODUCCIÓN

En función del análisis realizado en el capítulo 3 correspondiente a la Descripción del Problema, se estima de interés recordar el problema abierto que se aborda en este trabajo de tesis.

Al considerar las principales causas de fracasos para proyectos de Ingeniería de Software y de Explotación de Información, se pone de manifiesto la necesidad de proveer mecanismos que permitan al Ingeniero de Explotación de Información responder a ciertos interrogantes asociados a la gestión inicial del proyecto. Por una parte, es necesario establecer si es posible realizar el proyecto, si la solución es la correcta y si es posible cumplir con todas las metas y expectativas del mismo; es decir, si el mismo puede concluir de manera exitosa. Por otra parte, también es preciso determinar la cantidad de personas que van a participar en el equipo de trabajo durante el desarrollo del proyecto, y por cuanto tiempo.

En virtud de lo expuesto, el presente trabajo de tesis tiene como objetivo proponer, estudiar y validar dos modelos que, en función del análisis que se realice de las características de los proyectos de Explotación de Información, permitan dar solución a dichos problemas. Una vez identificados los requerimientos del proyecto, el primer modelo propuesto permite realizar la ***Evaluación de la Viabilidad*** del proyecto, a los efectos de determinar de manera temprana los puntos débiles asociados a aquellas características del proyecto que pueden presentar dificultades durante su desarrollo; mientras que el segundo modelo, tiene como objetivo realizar la ***Estimación del Esfuerzo*** (en tiempo/hombre) que se requiere para llevar a cabo el proyecto de manera satisfactoria y así poder planificar las actividades que se deben desarrollar. Asimismo, ambos modelos contemplan las particularidades de los proyectos de Explotación de Información realizados en el marco de las Pequeñas y Medianas Empresas (PyMEs); de esta forma, proporcionan respuestas a

los interrogantes antes mencionados en ese tipo de proyectos. La ilustración de esta idea se puede visualizar en la Figura 4.1.

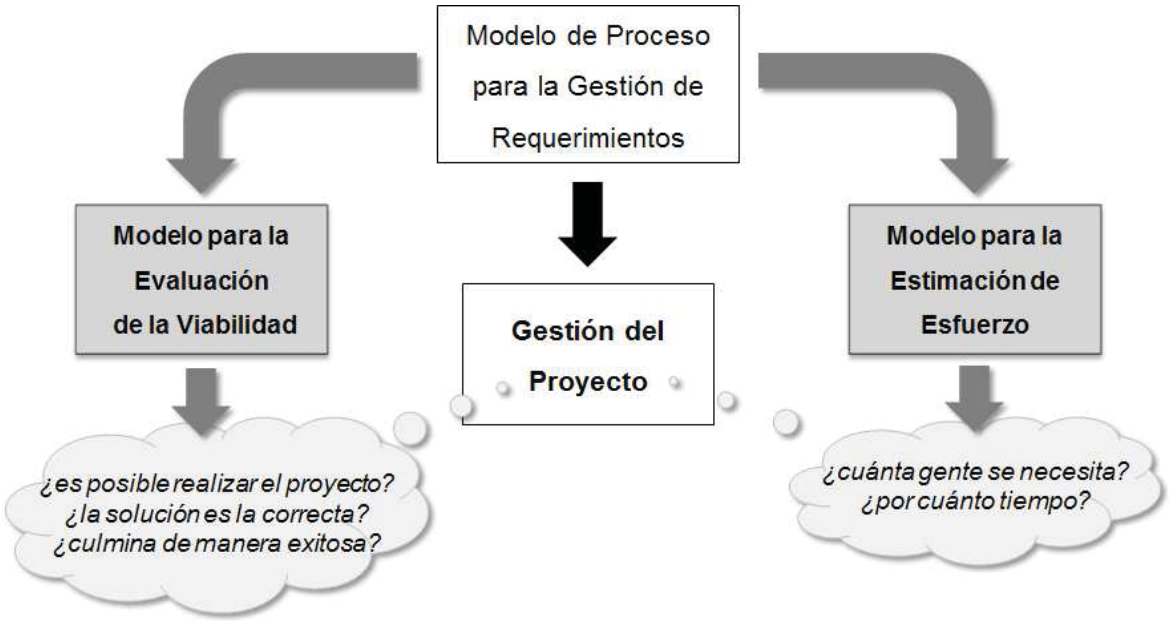


Fig. 4.1. Modelos Propuestos para responder los interrogantes asociados a la gestión inicial del proyecto.

Para definir ambos modelos se han consultado varias fuentes documentales que tratan las principales características de proyectos de Explotación de Información, los modelos de viabilidad existentes en la Ingeniería de Software y la Ingeniería del Conocimiento; así como también, los modelos de estimación en existencia para proyectos de Explotación de Información y de Ingeniería de Software.

Por otra parte, se han considerado como casos de estudio proyectos reales de Explotación de Información que han sido recolectados por investigadores de los siguientes grupos: Grupo de Investigación en Sistemas de Información del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús (GISI-DDPyT-UNLa), Grupo de Estudio en Metodologías de Ingeniería de Software de la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional (GEMIS-FRBA-UTN), y Grupo de Investigación en Explotación de Información en el Laboratorio de Informática Aplicada de la Universidad Nacional de Río Negro (GIEdI-UNRN). Cabe señalar que dichos proyectos fueron realizados aplicando la metodología CRISP-DM [Chapman *et al.*, 2000], por lo que los modelos propuestos se consideran confiables para proyectos de Explotación de Información que se desarrollen en base a dicha metodología. No obstante, y debido a que el Modelo de Proceso para Proyectos de Explotación de Información definido en [Vanrell *et al.*, 2010; 2012] se basa en la metodología CRISP-DM, los modelos propuestos también pueden ser aplicados en proyectos que apliquen dicho Modelo de Proceso.

En el caso de utilizar la metodología CRISP-DM pura, el Ingeniero encargado del proyecto debe aplicar los modelos propuestos en este trabajo de tesis dentro de la tarea general '*Evaluar la Situación*' correspondiente a la fase '*Comprensión del Negocio*'. En cambio, al utilizar el Modelo de Proceso para Proyectos de Explotación de Información, los modelos propuestos en esta tesis se deben ser aplicados en el contexto del subproceso '*Planificación / Entendimiento del Negocio*' correspondiente al proceso general de '*Administración de Proyectos*'.

4.2. MODELO PARA EVALUACIÓN DE LA VIABILIDAD DE PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

En esta sección se presenta la propuesta del modelo de evaluación correspondiente de la viabilidad de un proyecto de Explotación de Información dentro de una PyME, la cual se estructura en tres partes: generalidades del modelo (sección 4.2.1), propuesta del modelo (sección 4.2.2) y análisis preliminar del mismo (sección 4.2.3).

4.2.1. Generalidades del Modelo para Evaluación de la Viabilidad

Como ha sido mencionado anteriormente, el primer problema identificado en el capítulo 3 de este trabajo de tesis tiene que ver con la dificultad de poder anticipar al inicio del proyecto los principales problemas que puedan tener lugar. Esto genera que muchos proyectos sean cancelados antes de su finalización [Edelstein & Edelstein, 1997] o que los resultados obtenidos no sean de utilidad para la organización [Strand, 2000]. Por consiguiente, se considera necesario identificar estos puntos débiles en forma temprana, para que luego puedan ser controlados durante su desarrollo por las tareas de gestión de riesgos. A tal efecto, se propone este modelo que permite evaluar la viabilidad del proyecto a partir de las metas del proyecto, la relación de los interesados (stakeholders, en inglés), el tipo de tecnología a ser aplicada, y particularidades de los datos disponibles, entre otros aspectos.

El Ingeniero de Explotación de Información encargado del proyecto deberá aplicar este modelo durante la actividad '*Identificar Riesgos*' de la tarea específica '*Riesgos y contingencias*' correspondiente a la tarea general '*Evaluar la Situación*' en la fase '*Comprensión del Negocio*' de la metodología CRISP-DM. La ubicación de dicha tarea específica dentro de la metodología CRISP-DM se puede visualizar en la Figura 4.2. Si, en cambio, se utiliza el Modelo de Proceso, esa actividad es reemplazada por la tarea '*Evaluación de la situación*' del subproceso '*Planificación / Entendimiento del Negocio*' que se ubica en el proceso general de '*Administración de Proyectos*', como se puede observar en la Figura 4.3.

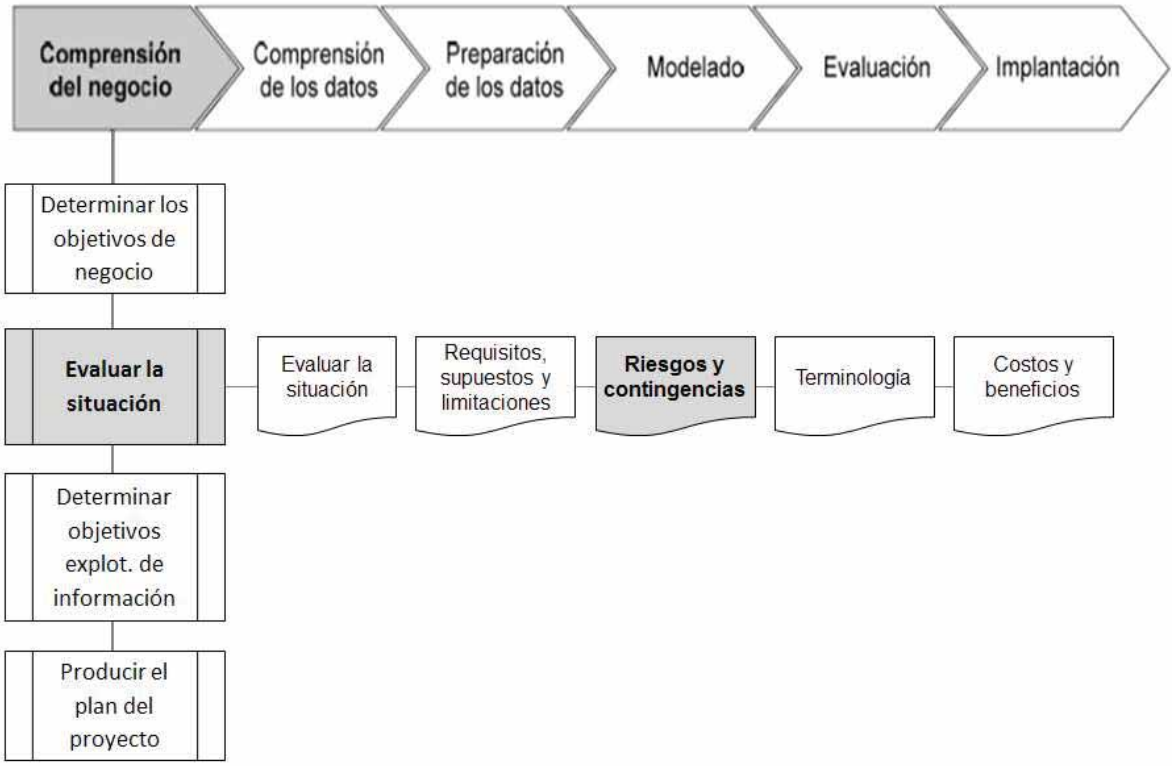


Fig. 4.2. Ubicación de la tarea de la metodología CRISP-DM en la que se aplica el Modelo de Viabilidad propuesto.

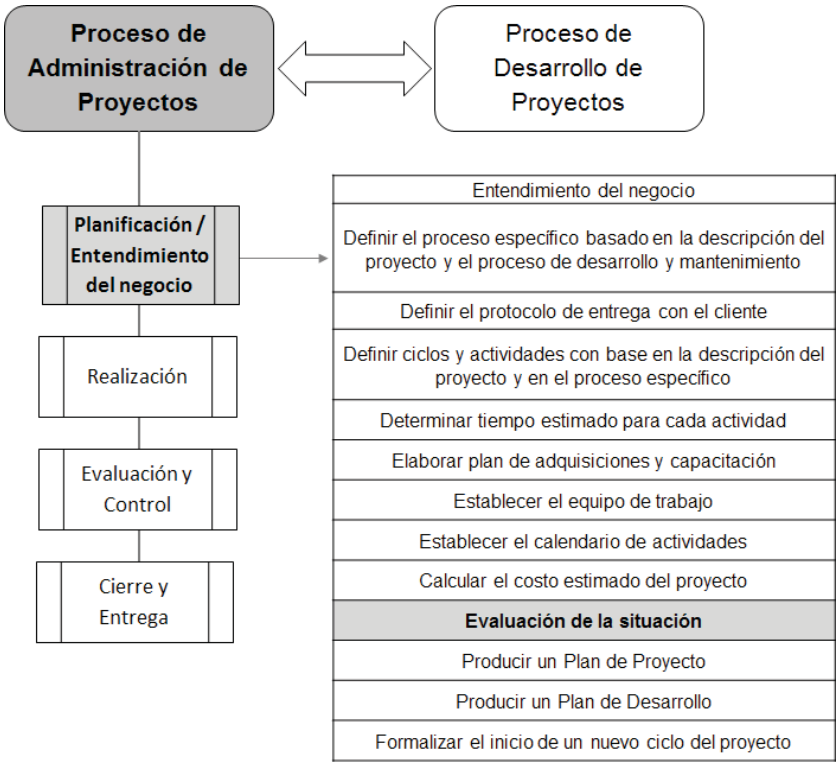


Fig. 4.3. Ubicación de la tarea del Modelo de Proceso en la que se aplica el Modelo de Viabilidad propuesto.

Para llevar a cabo la evaluación de la viabilidad, el modelo propuesto necesita que el Ingeniero de Explotación de Información responda a un conjunto de preguntas, cuyas respuestas permitan caracterizar al proyecto. Sin embargo, al comienzo de un proyecto no resulta sencillo contestar estas preguntas con un adecuado grado de certeza; como puede ser proporcionando respuesta del tipo '*sí*' / '*no*' o con un valor numérico. En virtud de estas consideraciones, el modelo propuesto se basa en el principio de los Sistemas Expertos Difusos [Jang, 1997], permitiendo de esta manera, manejar un rango de cinco valores lingüísticos (entre '*nada*' y '*todo*') y así dar respuesta a cada una de las preguntas considerar. Al hacer uso de un procedimiento sencillo es posible transformar los valores lingüísticos, indicados por el Ingeniero de Explotación de Información, en intervalos difusos que luego sean utilizados para obtener la valoración global de la viabilidad del proyecto. Asimismo, esta valoración se analiza teniendo en cuenta tres grupos o dimensiones del proyecto en forma similar a los utilizados por el Test de Viabilidad para proyectos en Ingeniería del Conocimiento (INCO) descripto en [García-Martínez & Britos, 2004; Gómez *et al.*, 1997; López *et al.*, 1991]. Las tres dimensiones a considerar en un proyecto de Explotación de Información son:

- ***Plausibilidad del proyecto:*** Esta dimensión incluye todas las características que hacen posible realizar el proyecto de explotación de información.
- ***Adecuación del proyecto:*** Incluye todas las características que determinan que la explotación de información es la solución apropiada para el problema de negocio detectado (es decir es la mejor solución para el problema).
- ***Éxito del proyecto:*** Incluye todas aquellas características que aseguran el éxito del proyecto de explotación de información.

De esta forma, un proyecto puede ser considerado como *viable* si el mismo es *plausible* de ser desarrollado, es capaz de proporcionar la solución *apropiada* para el problema que le dio origen, y tiene aceptables posibilidades de alcanzar el *éxito*. Como se puede ver, la primera cuestión analiza si se cumplen las condiciones necesarias para poder desarrollar el proyecto, mientras que las otras dos predicen si las expectativas del proyecto podrán ser satisfechas mediante la aplicación de la tecnología seleccionada.

Cabe destacar, que en este análisis no se ha incluido la evaluación de la dimensión *justificación* del proyecto, que sí es considerada en proyectos de la INCO. Se recuerda que dichos proyectos tienen como objetivo construir un sistema software para emular parte del comportamiento de un experto humano en la realización de una tarea. Por lo tanto, dentro del Test de Viabilidad se debe considerar si vale la pena desarrollar el sistema software teniendo en cuenta la posibilidad de perder los

conocimientos del experto. Por el contrario, de proyectos de Explotación de Información no buscan reemplazar a los expertos existentes en la organización, sino que intenta asistir a la toma de decisiones mediante el análisis de los datos disponibles en la organización. Como, además, no existe la posibilidad de perder dichos datos, no se ha considerado necesario utilizar toda una dimensión para evaluar la justificación del proyecto. A tal efecto, las características que se podrían relacionar a la misma, se encuentran incluidas dentro de las otras tres dimensiones analizadas por el modelo (Plausibilidad, Adecuación y Éxito).

4.2.2. Propuesta del Modelo para la Evaluación de la Viabilidad

Un modelo permite identificar, definir e integrar distintos elementos de una realidad para ayudar su análisis. Para poder proponer este modelo, primero es necesario identificar las principales condiciones que un proyecto de Explotación de Información debe cumplir para ser considerado como viable (indicadas en la sección 4.2.2.1). Con las condiciones ya identificadas se define el proceso para realizar el cálculo del valor de cada dimensión y la viabilidad global del proyecto. Este proceso consta de cinco pasos que se indican a continuación:

1. Determinar el valor correspondiente para cada una de las características del proyecto.
2. Convertir los valores en intervalos difusos.
3. Calcular la valoración de cada dimensión.
4. Calcular la valoración global de la viabilidad del proyecto.
5. Interpretar los resultados obtenidos.

Cada paso de este proceso se encuentra detallado en la sección 4.2.2.2.

4.2.2.1. Condiciones a considerar para la Evaluación de la Viabilidad

A partir de la investigación documental realizada en [Bolea *et al.*, 2011; Davenport, 2009; Fayyad, 2000; Lavravc *et al.*, 2004; Nadali *et al.*, 2011; Nemati & Barkom, 2003; Nie *et al.*, 2009; Pipino *et al.*, 2002; Sim, 2003] se determinan los conceptos de interés para evaluar la viabilidad de los proyectos considerados. Estas condiciones han sido clasificadas en las tres dimensiones indicadas en la sección 4.2.1 en forma similar al criterio empleado por el Test de Viabilidad para proyectos de la INCO.

Los tres grupos de condiciones evaluadas son:

- ***Condiciones que determinan la plausibilidad del proyecto:***

  Un proyecto puede ser realizado con explotación de información si se cumplen las siguientes condiciones:

  - los repositorios disponibles poseen datos actuales y representativos para el problema de negocio que se desea solucionar;
  - se puede entender claramente el problema de negocio a resolver; y
  - existen personas en el equipo de trabajo con un mínimo conocimiento sobre explotación de información.

- ***Condiciones que determinan la adecuación del proyecto:***

  Es adecuado aplicar explotación de información en un proyecto si se cumplen las siguientes condiciones:

  - los repositorios disponibles se encuentran en formato digital (es decir, no sólo se pueden acceder a los datos en papel impreso);
  - las técnicas estadísticas tradicionales no permiten encontrar una buena solución al problema de negocio;
  - el problema de negocio no es muy variable durante el desarrollo del proyecto; y
  - la calidad de los datos es buena.

    En caso de que la calidad de los datos no sea buena, los resultados de la minería de datos tampoco lo serán [Han & Kamber, 2011]. Para evaluar la calidad de los datos se utilizarán las siguientes métricas:

    - Cantidad de atributos y registros (mide que se disponga de una cantidad suficiente de datos para aplicar minería de datos).
    - Grado de credibilidad de los datos (mide cuanto se puede confiar que los datos son verdaderos dependiendo de su fuente y naturaleza).

- ***Condiciones que determinan el éxito del proyecto:***

  Un proyecto desarrollado con explotación de información será exitoso si se cumplen las siguientes condiciones:

  - los repositorios de datos están implementados con tecnologías que permiten un fácil acceso y manipulación (tareas de integración, limpieza y formateo);
  - se cuenta con el apoyo de los principales interesados en el proyecto (stakeholders) que pueden ser tanto los directivos de la organización, los gerentes de medio nivel y/o los usuarios finales;
  - existen personas en el equipo de trabajo con experiencia en proyectos similares; y

- es posible realizar una planificación completa, correcta y verdadera del proyecto considerando la realización de buenas prácticas ingenieriles con el tiempo adecuado.

4.2.2.2. Proceso para Evaluación de la Viabilidad

A continuación se describen los cinco pasos que se deben realizar para evaluar la viabilidad de un proyecto de explotación de información:

Paso 1: Determinar el valor correspondiente para cada una de las características del proyecto.

Para caracterizar un proyecto de Explotación de Información, y evaluar luego su viabilidad se utilizan las características definidas en la Tabla 4.1 las cuales están basadas en las condiciones indicadas en la sección 4.2.2.1. A partir del resultado de las entrevistas realizadas en la organización, el ingeniero debe responder las preguntas asociadas a cada característica. Los valores lingüísticos permitidos para las respuestas son '*nada*', '*poco*', '*regular*', '*mucho*' y '*todo*', donde cuanto más verdadera parezca una característica, mayor valor se le debe asignar; y, cuanto más falsa parezca, menor valor.

Categoría	ID	Pregunta asociada a la Característica	Peso	Umbral
Datos	P1	¿En qué medida los repositorios disponibles poseen datos actuales?	*8*	*poco*
	P2	¿Qué tan representativos son los datos de los repositorios disponibles para resolver el problema de negocio?	*9*	*poco*
	A1	¿En qué medida los repositorios se encuentran disponibles en formato digital?	*4*	*poco*
	A2	¿Qué cantidad de atributos y registros tienen los datos disponibles?	*7*	*poco*
	A3	¿Cuánta confianza se posee en la credibilidad de los datos disponibles?	*8*	*poco*
	E1	¿Cuánto facilita la tecnología de los repositorios disponibles las tareas de manipulación de los datos?	*6*	*nada*
Problema de Negocio	P3	¿Cuánto se entiende del problema de negocio?	*7*	*poco*
	A4	¿En qué medida el problema de negocio no puede ser resuelto aplicando técnicas estadísticas tradicionales?	*10*	*poco*
	A5	¿Qué tan estable es el problema de negocio durante el desarrollo del proyecto?	*9*	*poco*
Tipo de Proyecto	E2	¿Cuánto apoyan los interesados (stakeholders) al proyecto?	*8*	*nada*
	E3	¿En qué medida la planificación del proyecto permite considerar la realización de buenas prácticas ingenieriles con el tiempo adecuado?	*7*	*nada*
Equipo de Trabajo	P4	¿Qué nivel de conocimientos posee el equipo de trabajo sobre explotación de información?	*6*	*poco*
	E4	¿Qué nivel de experiencia posee el equipo de trabajo en proyectos similares?	*6*	*nada*

Tabla 4.1. Características a ser evaluadas por el modelo de viabilidad.

Para cada característica de la tabla se definen los siguientes atributos:

- *Categoría* que es utilizado únicamente para poder agrupar las características de acuerdo a qué o quién se refiere. Las categorías analizadas son los *Datos* disponibles para ser utilizados, el *Problema de Negocio* que se intenta resolver, y cuestiones particulares del *Tipo de Proyecto* y el *Equipo de Trabajo* que va a participar en el mismo.
- *ID* que indica el código para identificar unívocamente a la característica y a la dimensión a la que pertenece.
- *Pregunta asociada a la Característica* que describe la condición que debe responder el ingeniero con el valor lingüístico correspondiente.
- *Peso* que indica la importancia relativa a cada característica en la globalidad del modelo. Como se puede ver la suma de todos los pesos no es igual a 100, pero esto es soportado por las fórmulas utilizadas en el modelo.
- *Umbral* que define el valor que la característica debe igualar o superar. En caso de que el valor asignado a la característica no supere el umbral, se puede considerar que el proyecto no es viable y no es necesario continuar con los pasos siguientes.

Paso 2: Convertir los valores en intervalos difusos.

Una vez que para cada característica de la Tabla 4.1 se han asignado los valores lingüísticos correspondientes, estos valores se deben traducir en intervalos difusos. En este sentido, a cada valor lingüístico se le define un intervalo difuso expresado por cuatro valores numéricos (entre cero y diez) que representan los puntos de ruptura (o puntos angulares) de su función de pertenencia correspondiente.

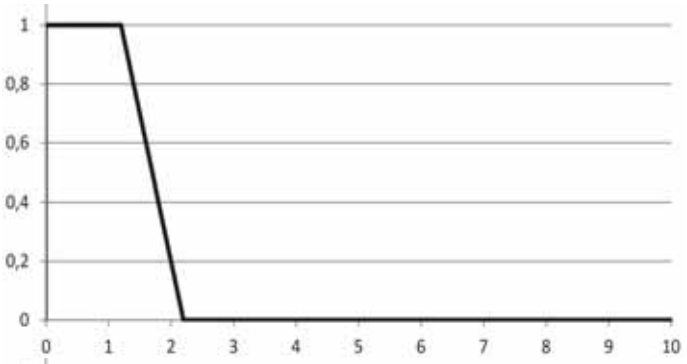
Estos intervalos, junto con la representación gráfica de la función de pertenencia, se indican en la Figura 4.4.

Paso 3: Calcular la valoración de cada dimensión.

Una vez obtenidos los intervalos difusos en el paso anterior, los mismos son utilizados para calcular la valoración de cada dimensión del proyecto.

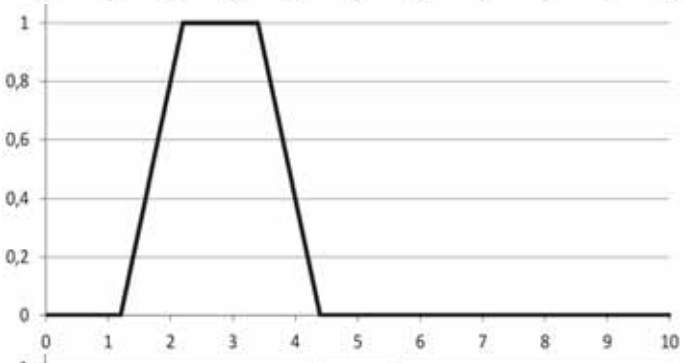
Los intervalos son agrupados por dimensión para ser ponderados considerando su peso correspondiente (los cuales fueron indicados en la Tabla 4.1). Esto se realiza aplicando una fórmula formada por la combinación de la media armónica y la media aritmética del conjunto de intervalos. De esta forma se busca reducir la influencia de valores bajos en el cálculo de la dimensión. Como resultado se obtiene un intervalo que representa la valoración de cada dimensión (I_d).

Valor = '*nada*'



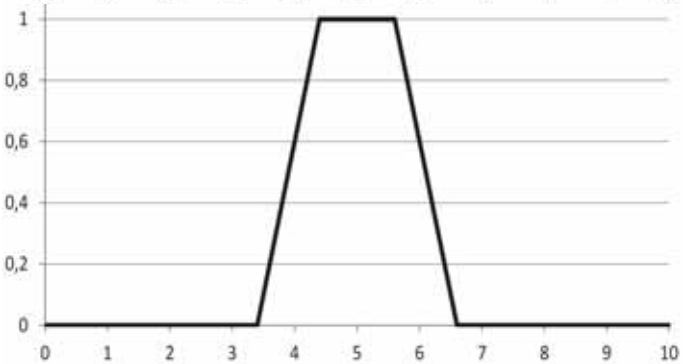
Intervalo Difuso = (0,01; 0,01; 1,20; 2,20)

Valor = '*poco*'



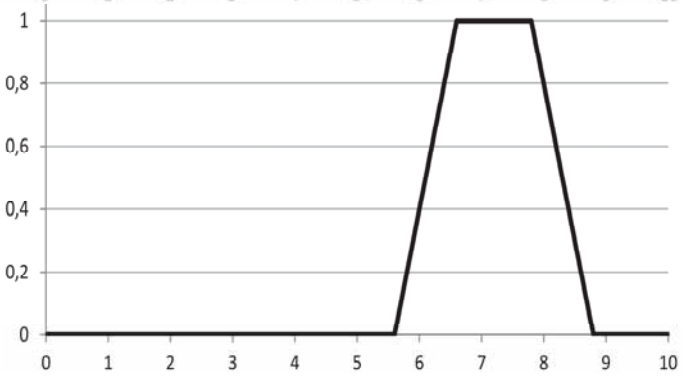
Intervalo Difuso = (1,20; 2,20; 3,40; 4,40)

Valor = '*regular*'



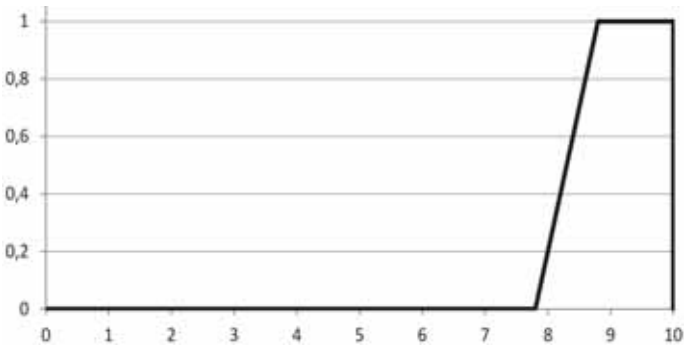
Intervalo Difuso = (3,40; 4,40; 5,60; 6,60)

Valor = '*mucho*'



Intervalo Difuso = (5,60; 6,60; 7,80; 8,80)

Valor = '*todo*'



Intervalo Difuso = (7,80; 8,80; 10,0; 10,0)

Fig. 4.4. Representación de la Función de Pertenencia y asignación de Intervalo Difuso para los Valores Lingüísticos.

La fórmula para calcular el intervalo de cada dimensión es la siguiente:

$$I_d = \left(\frac{1}{2} \cdot \frac{\sum_{i=1}^{n_d} P_{d_i}}{\sum_{i=1}^{n_d} \left(\frac{P_{d_i}}{C_{d_i}} \right)} \right) + \left(\frac{1}{2} \cdot \frac{\sum_{i=1}^{n_d} (P_{d_i} \cdot C_{d_i})}{\sum_{i=1}^{n_d} P_{d_i}} \right)$$

Donde:

I_d: representa el intervalo difuso calculado para la dimensión *d* (usando como nomenclatura '*P*' para plausibilidad, '*A*' para adecuación y '*E*' para criterio de éxito).

P_{di}: representa el peso de la característica *i* perteneciente a la dimensión *d*.

C_{di}: representa el intervalo difuso asignado a la característica *i* perteneciente a la dimensión *d*.

n_d: representa la cantidad de características asociada a la dimensión *d*.

Dado que el resultado de la fórmula anterior es otro intervalo difuso, para convertir dicho intervalo en un único valor numérico (V_d) se utiliza la media aritmética de los valores del intervalo como se indica en la siguiente fórmula:

$$V_d = \frac{\sum_{j=1}^{4} (I_d\{j\})}{4}$$

Donde:

V_d: representa el valor numérico calculado para la dimensión *d*.

$I_d\{j\}$: representa el valor correspondiente a la posición *j* del intervalo difuso calculado para la dimensión *d*.

<u>Paso 4:</u> Calcular la valoración global de la viabilidad del proyecto.

Finalmente, los valores numéricos calculados en el paso anterior para cada dimensión (V_d) son combinados a través de una media aritmética ponderada y así se consigue el valor de la viabilidad global del proyecto (EV).

La fórmula propuesta es la siguiente:

$$EV = \frac{8 \cdot V_P + 8 \cdot V_A + 6 \cdot V_E}{22}$$

Donde:

EV: representa el valor global de la viabilidad del proyecto.
V_P*:* representa el valor para la dimensión plausibilidad.
V_A*:* representa el valor para la dimensión adecuación.
V_E*:* representa el valor para la dimensión criterio de éxito.

Paso 5: Interpretar los resultados obtenidos.

Una vez que los valores correspondientes a cada dimensión y al proyecto global son calculados (pasos 3 y 4 respectivamente), deben ser analizados e interpretados.

Primero se determina los resultados de la viabilidad de cada dimensión, los cuales se recomienda graficar utilizando la función de pertenencia correspondiente a su intervalo difuso (I_d). Se considera que una dimensión está aceptada si el gráfico obtenido supera al intervalo del valor '*regular*'. En forma análoga esto se puede determinar si se utiliza el valor numérico de la dimensión: una dimensión será aceptable si su valor V_d es mayor a 5.

Luego, para la viabilidad global del proyecto se utiliza el siguiente criterio: si las tres dimensiones son aceptadas (de acuerdo a la regla mencionada anteriormente) y la valoración global de la viabilidad proyecto (EV) es mayor a 5 entonces el proyecto es viable. En caso contrario, el proyecto no es viable.

En ambos casos, el ingeniero deberá observar los puntos débiles del proyecto que deben ser reforzados (en caso de proyecto no viable) y/o deben ser monitoreados durante el desarrollo del proyecto (si el proyecto fue encontrado como viable).

4.2.3. Análisis Preliminar del Modelo para Evaluación de la Viabilidad

En esta sección se realiza un análisis preliminar del modelo propuesto (un análisis más completo para su validación se describe en el capítulo siguiente). Este análisis preliminar intenta ilustrar el uso del modelo mediante una prueba de concepto (sección 4.2.3.1); así como también realizar un análisis estadístico del comportamiento del modelo por medio del método Monte Carlo, el cual se explica con mayor detalle en la sección 4.2.3.2.

4.2.3.1. Prueba de Concepto del Modelo para Evaluación de la Viabilidad

En esta sección se presentan dos pruebas de concepto para ilustrar el funcionamiento del modelo que se propone en este trabajo. Se utiliza una prueba de concepto positiva con un proyecto de Explotación de Información real finalizado con éxito y, por consiguiente, viable; y otra negativa, modificando algunos datos de dicho proyecto para que el mismo no sea viable. Todos los cálculos necesarios fueron realizados mediante una planilla creada ad-hoc [Pytel, 2012a] que implementa las fórmulas definidas anteriormente.

El proyecto evaluado tenía como objetivo identificar las evidencias de causalidad entre la satisfacción general, el servicio contratado y la baja de clientes de los clientes de una organización proveedora de Internet; a tal efecto, se hizo uso de la información recolectada de una encuesta realizada a sus clientes por parte de la organización.

Para llevar a cabo la prueba de concepto positiva se aplica el proceso propuesto en la sección 4.2.2.2. A partir de diversas sesiones de educción en la organización, se definió el problema de negocio, las principales características de la organización y los datos disponibles. Con esta información se respondieron las preguntas requeridas con el valor lingüístico correspondiente (paso 1), que se muestran en la Tabla 4.2. Estos valores son convertidos en intervalos difusos (paso 2), para luego calcular el intervalo de cada dimensión (paso 3) y se representa en forma gráfica como se muestra en la Tabla 4.3. Por último, se calcula el valor numérico de cada dimensión y la valoración global de la viabilidad del proyecto (paso 4), los cuales se interpretan (paso 5) como se indica en la Tabla 4.4.

Para realizar la prueba de concepto negativa, se modificaron los valores de las seis características asociadas a los *datos* y al *problema de negocio* de la Tabla 4.2, con los valores '*regular*' (para las características P1, P3 y A4) y '*poco*' (para A1, A2 y A5). Estos nuevos valores no son inferiores a los umbrales para no generar un ejemplo negativo trivial; esto se genera cuando un proyecto posee alguna característica que no supera el valor del umbral. De igual manera a la prueba de concepto positiva, se lleva a cabo el cálculo de los intervalos y valores correspondientes; los resultados obtenidos se muestran en las Tablas 4.5 y 4.6. Haciéndose notar que en la primera tabla, la dimensión Éxito no se muestra ya que sus resultados no se modifican con respecto a los indicados en la Tabla 4.2.

Categoría	ID	Respuesta	Valor Asig.
Datos	P1	Los repositorios poseen datos contestados por clientes actuales de la organización.	*todo*
	P2	No se posee información de todos los clientes que se han dado de baja en los últimos 6 meses.	*regular*
	A1	Las respuestas de la encuesta se encuentran totalmente digitalizadas.	*todo*
	A2	Se cuenta con apropiadamente 65.000 registros y 20 atributos para ser utilizados.	*mucho*
	A3	La encuesta ha sido respondida por los clientes sin supervisión a través de una aplicación web.	*regular*
	E1	La encuesta digitalizada ha sido suministrada en un archivo de texto para ser procesada.	*poco*
Problema de Negocio	P3	Varias sesiones de educción se han realizado para determinar el objetivo del proyecto para solucionar el problema de negocio detectado.	*todo*
	A4	La organización no cuenta con ningún experto disponible en los datos a ser utilizados que pueda definir alguna hipótesis para ser probadas por técnicas estadísticas. Se considera que la "mejor solución" es la aplicación de técnicas de minería de datos.	*mucho*
	A5	No se posee mucha seguridad que el problema de negocio detectado se mantenga durante todo el proyecto ya que depende del comportamiento de los clientes.	*regular*
Tipo del Proyecto	E2	El gerente de sistemas y el de marketing tienen grandes intereses en la finalización exitosa de este proyecto.	*mucho*
	E3	La organización desea recibir los resultados en el menor tiempo posible.	*regular*
Equipo de Trabajo	P4	El equipo de trabajo posee gran conocimientos sobre explotación de información en general y las técnicas de minería de datos en particular.	*todo*
	E4	El equipo de trabajo posee experiencia en la aplicación de explotación de información en proyectos similares pero con diferentes datos.	*mucho*

Tabla 4.2. Asignación de las características del proyecto utilizado como prueba positiva.

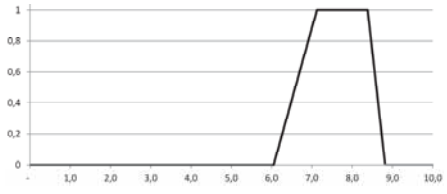
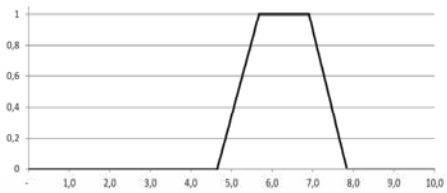
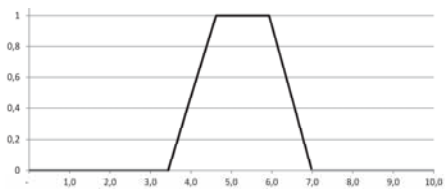
Dimensión	ID	Intervalo Difuso del Valor Asignado	Intervalo Valor Dimensión (I_d)	Representación de la Función de Pertenencia de I_d
Plausibilidad	P1	(7,8; 8,8; 10; 10)	(6,05; 7,12; 8,39; 8,82) Intervalo sobrepasa por pequeña diferencia al valor '*mucho*'.	
	P2	(3,4; 4,4; 5,6; 6,6)		
	P3	(7,8; 8,8; 10; 10)		
	P4	(7,8; 8,8; 10,0; 10,0)		
Adecuación	A1	(7,8; 8,8; 10,0; 10,0)	(4,65; 5,68; 6,91; 7,84) Intervalo entre valores '*regular*' y '*mucho*'.	
	A2	(5,6; 6,6; 7,8; 8,8)		
	A3	(3,4; 4,4; 5,6; 6,6)		
	A4	(5,6; 6,6; 7,8; 8,8)		
	A5	(3,4; 4,4; 5,6; 6,6)		
Éxito	E1	(1,2; 2,2; 3,4; 4,4)	(3,44; 4,62; 5,93; 6,99) Intervalo sobrepasa por pequeña diferencia al valor '*regular*'	
	E2	(5,6; 6,6; 7,8; 8,8)		
	E3	(3,4; 4,4; 5,6; 6,6)		
	E4	(5,6; 6,6; 7,8; 8,8)		

Tabla 4.3. Traducción y cálculo de intervalos por dimensión para prueba de concepto positiva.

Dimensión	Valor de la Dimensión (V_d)	Interpretación
Plausibilidad	*7,60*	Dado que los tres valores son superiores al valor mínimo requerido de 5, la viabilidad de todas las dimensiones está aceptada. Sin embargo, debe notarse que a pesar de que la valoración de Plausibilidad y Adecuación es holgada, para el Éxito del proyecto es muy cercana al valor mínimo requerido. Esto significa que durante el proyecto habrá que monitorear con mayor atención a las características evaluadas para el éxito.
Adecuación	*6,27*	
Éxito	*5,25*	
Valor global de la viabilidad del proyecto (*EV*)	***6,47***	***Se considera que el proyecto es viable para ser realizado.***

Tabla 4.4. Cálculo por dimensión y viabilidad global para la prueba de concepto positiva.

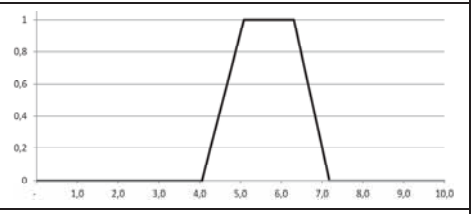
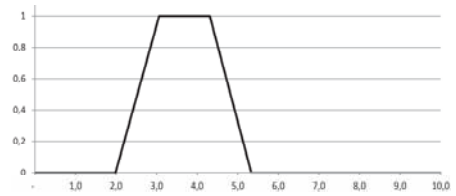
Dimensión	ID	Intervalo Difuso del Valor Asignado	Intervalo Valor Dimensión (I_d)	Representación de la Función de Pertenencia de I_d
Plausibilidad	P1	(3,4; 4,4; 5,6; 6,6)	(4,06; 5,08; 6,31; 7,18) Intervalo sobrepasa por pequeña diferencia al valor '*regular*'.	
	P2	(3,4; 4,4; 5,6; 6,6)		
	P3	(3,4; 4,4; 5,6; 6,6)		
	P4	(7,8; 8,8; 10,0; 10,0)		
Adecuación	A1	(1,2; 2,2; 3,4; 4,4)	(1,99; 3,06; 4,31; 5,34) Intervalo sobrepasa por pequeña diferencia al valor '*poco*'.	
	A2	(1,2; 2,2; 3,4; 4,4)		
	A3	(3,4; 4,4; 5,6; 6,6)		
	A4	(3,4; 4,4; 5,6; 6,6)		
	A5	(1,2; 2,2; 3,4; 4,4)		

Tabla 4.5. Traducción y cálculo de intervalos por dimensión para prueba de concepto negativa.

Dimensión	Valor de la Dimensión (V_d)	Interpretación
Plausibilidad	*5,66*	Dado que el valor para la Adecuación no supera el valor mínimo de 5, no se considera que la explotación de información sea la solución adecuada para este proyecto. Esto significa también que el proyecto no es viable para ser realizado. Por otro lado, los valores para Plausibilidad y el Éxito, a pesar de superar el mínimo, no lo hacen por mucho. Esto significa que los riesgos de realizar el proyecto serían altos.
Adecuación	*3,67*	
Éxito	*5,25*	
Valor global de la viabilidad del proyecto (*EV*)	***4,82***	***Se considera que el proyecto NO es viable para ser realizado.***

Tabla 4.6. Cálculo por dimensión y viabilidad global para la prueba de concepto negativa.

4.2.3.2. Análisis Estadístico del Modelo para Evaluación de la Viabilidad

Para realizar el análisis estadístico del comportamiento del modelo propuesto se ha utilizado el método de simulación Monte Carlo [Metropolis & Ulam, 1949]. Se han generado en forma pseudo-aleatoria un banco de pruebas con los datos de 25.000 proyectos cuyos datos se encuentran disponibles en [Pytel, 2012b]. A partir de los datos de los proyectos simulados, lo cuales presentan con diferentes valores para sus características, se han aplicado las fórmulas propuestas con el objetivo de calcular el valor correspondiente para cada dimensión y el valor global de la viabilidad.
Este análisis estadístico se divide en tres partes de acuerdo a las partes de la estructura del modelo presentado: por cada dimensión (sección 4.2.3.2.1), por cada categoría (sección 4.2.3.2.2) y la viabilidad global (sección 4.2.3.2.3). Finalmente, se presentan unas conclusiones preliminares del comportamiento del modelo a partir de estos estudios (sección 4.2.3.2.4).

4.2.3.2.1. Análisis Estadístico por Dimensión

El primer análisis consiste en la interpretación de los gráficos que representan la valoración de cada dimensión con respecto al valor de cada una de sus características asociadas, tal como se muestra en la Figuras 4.5, 4.6 y 4.7. Al observar estos gráficos, se puede ver que cuanto mayor valor adquieren las características, la valoración de la dimensión es mayor. Esto concuerda con la definición de las características presentada en la Sección 4.2.2.
A continuación se describe el análisis realizado para cada dimensión:

- En el caso de la *Plausibilidad* (Figura 4.5), el mínimo requerido para aceptar la dimensión se consigue cuando las características toman un valor superior a '*poco*'; siendo la característica más relevante el grado de representatividad de los datos (P2) por tener la mayor pendiente de crecimiento, tal como se puede observar en el gráfico donde esta característica toma los valores más extremos.

- Para aceptar la *Adecuación* (Figura 4.6), las características también deben superar el valor '*poco*'. No obstante, esto no es necesario para aquella característica que mide si los repositorios se encuentran en formato digital (A1); dado que se desprende de la figura que ésta no ejerce una influencia significativa sobre el grado de aceptación de la dimensión. Esto se debe a que para el menor valor de esta característica, la dimensión supera el mínimo requerido. Para esta dimensión, la principal característica detectada es la que evalúa si el problema de negocio puede resolverse aplicando técnicas estadísticas tradicionales (A4).

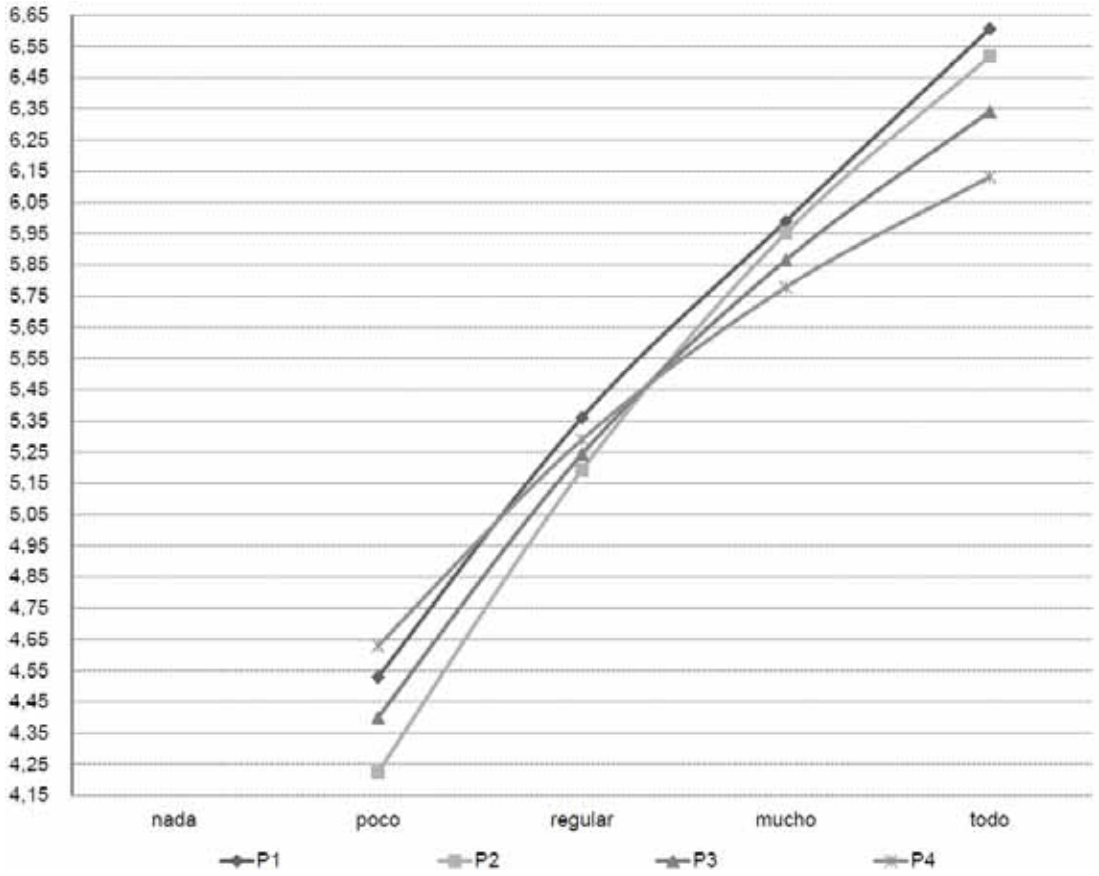


Fig. 4.5. Representación de la variación de las características para la Dimensión Plausibilidad.

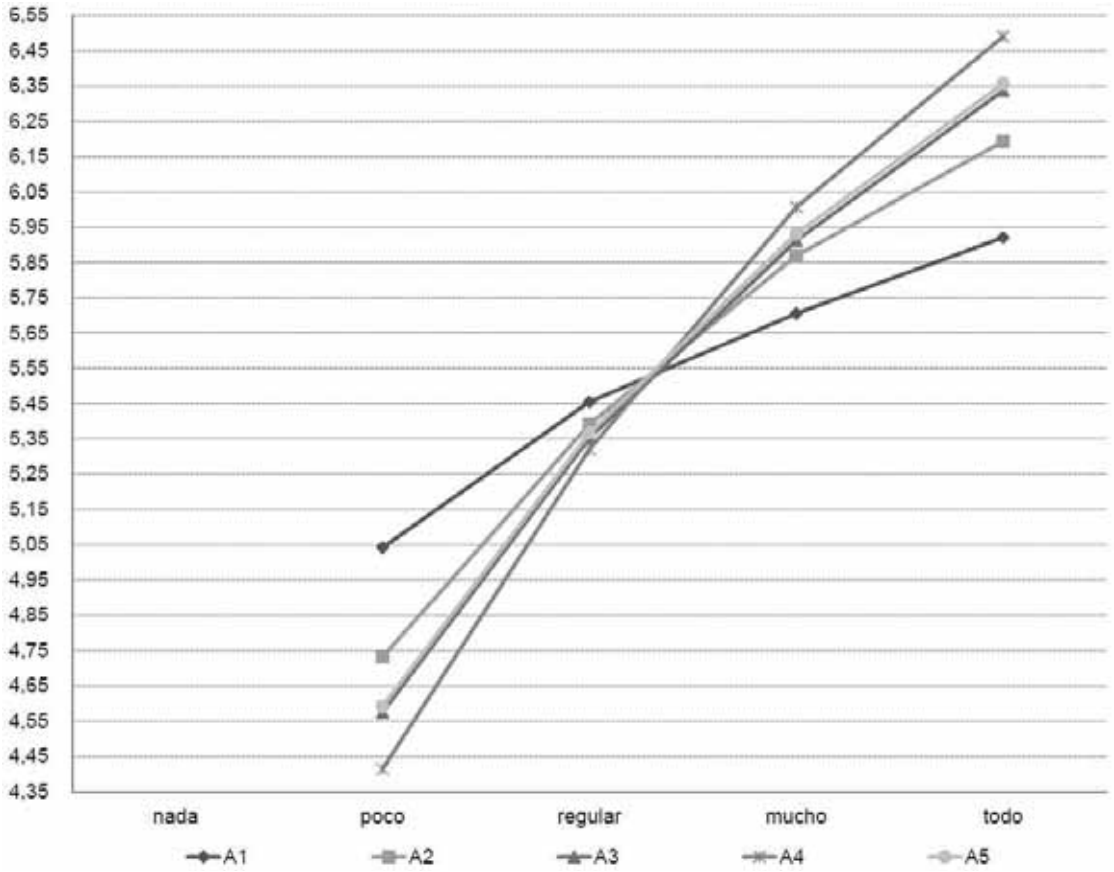


Fig. 4.6. Representación de la variación de las características para la Dimensión Adecuación.

- Por último, en el caso del *Éxito* (Figura 4.7) se observa que para aceptar la dimensión todas sus características deben tomar un valor superior a '*mucho*' (es decir, tendiendo a '*todo*'). Como todas las pendientes son similares, no se puede identificar ninguna característica principal para esta dimensión.

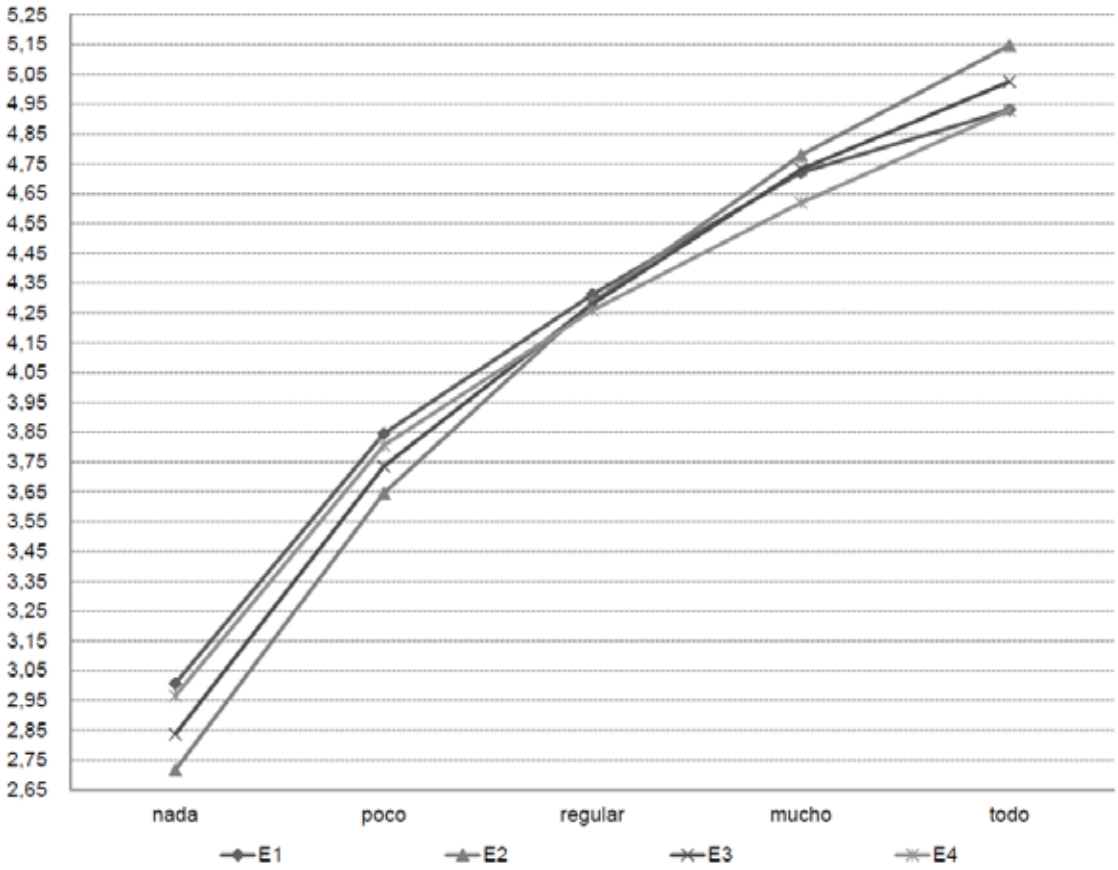


Fig. 4.7. Representación de la variación de las características para la Dimensión Éxito.

4.2.3.2.2. Análisis Estadístico por Categoría

Luego del análisis realizado en la sección anterior, se han generado gráficos similares pero en este caso se representa la variación de la viabilidad global del proyecto con respecto a las características agrupadas por la categoría a la que pertenecen (Figuras 4.8, 4.9, 4.10 y 4.11). A partir del análisis realizado sobre dichas figuras, se puede concluir que para que un proyecto de Explotación de Información sea viable se debe cumplir las siguientes condiciones:

- las características relacionadas con los *datos* (Figura 4.8) y el *problema de negocio* (Figura 4.9) deben tomar un valor mayor a '*poco*',

- las características relacionadas al *tipo de proyecto* (Figura 4.10) deben tener valores mayores a '*nada*'
- y, en el caso del *equipo de trabajo* (Figura 4.11), depende del valor de sus dos características: el valor P4 debe ser mayor a '*poco*' y el de E4 mayor a '*nada*'.

Esto confirma lo que intuitivamente se suponía: en proyectos de Explotación de Información realizados en el marco de una PyME, es importante obtener mejores valores de las características asociadas a los *datos* y el *problema de negocio* para que el proyecto sea viable; en cambio, las características menos importantes parecen ser las asociadas al *equipo de trabajo*.

De manera alternativa, es importante resaltar que con respecto a los *datos* (Figura 4.8), la principal característica para lograr que el proyecto sea viable es el grado de representatividad de los datos (P2). Otra característica que también se destaca es la que evalúa si la tecnología de los repositorios facilita la manipulación de los datos (E1), dado que esta posee un crecimiento más pronunciado. Finalmente, en las categorías asociadas al *problema de negocio*, al *tipo de proyecto* y al *equipo de trabajo*, no se identifica ninguna característica que se destaque por presentar todas las características una curva similar en las figuras 4.9, 4.10 y 4.11.

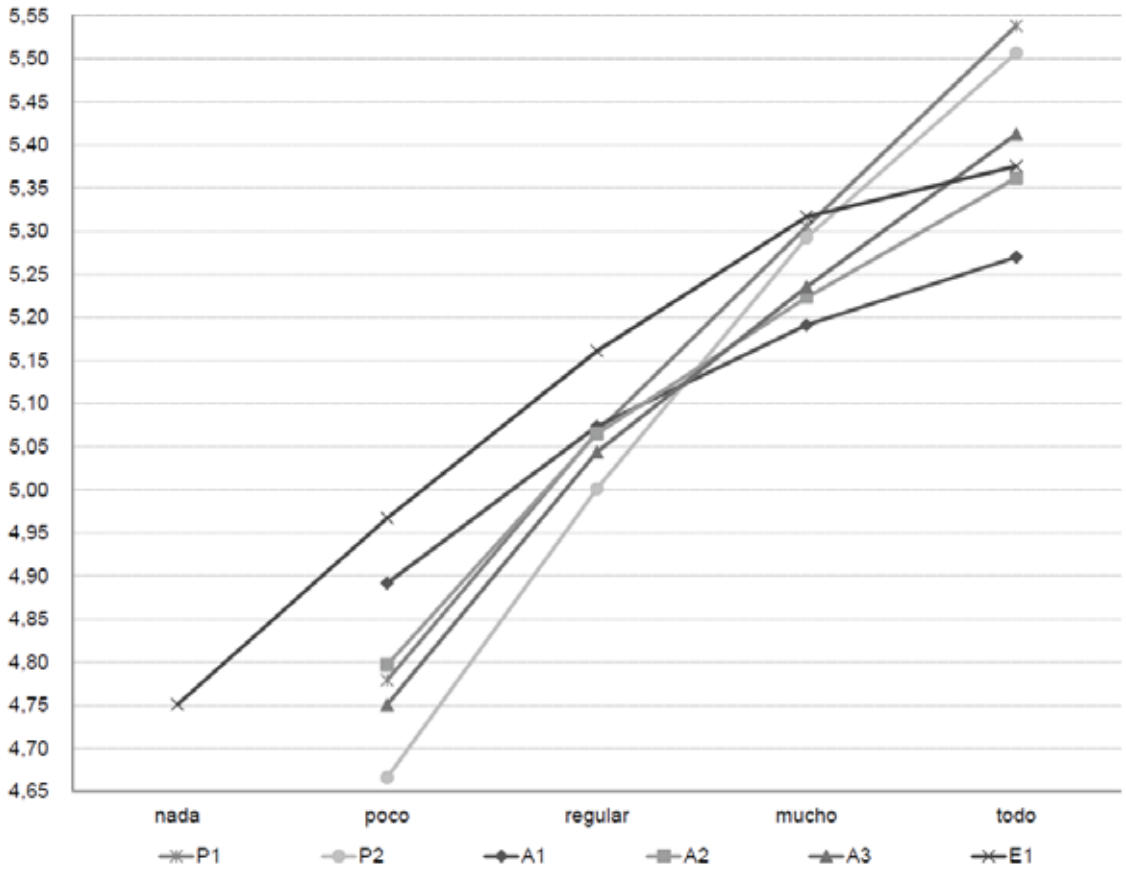


Fig. 4.8. Representación de la Viabilidad Global al cambiar el valor de las características de la categoría Datos.

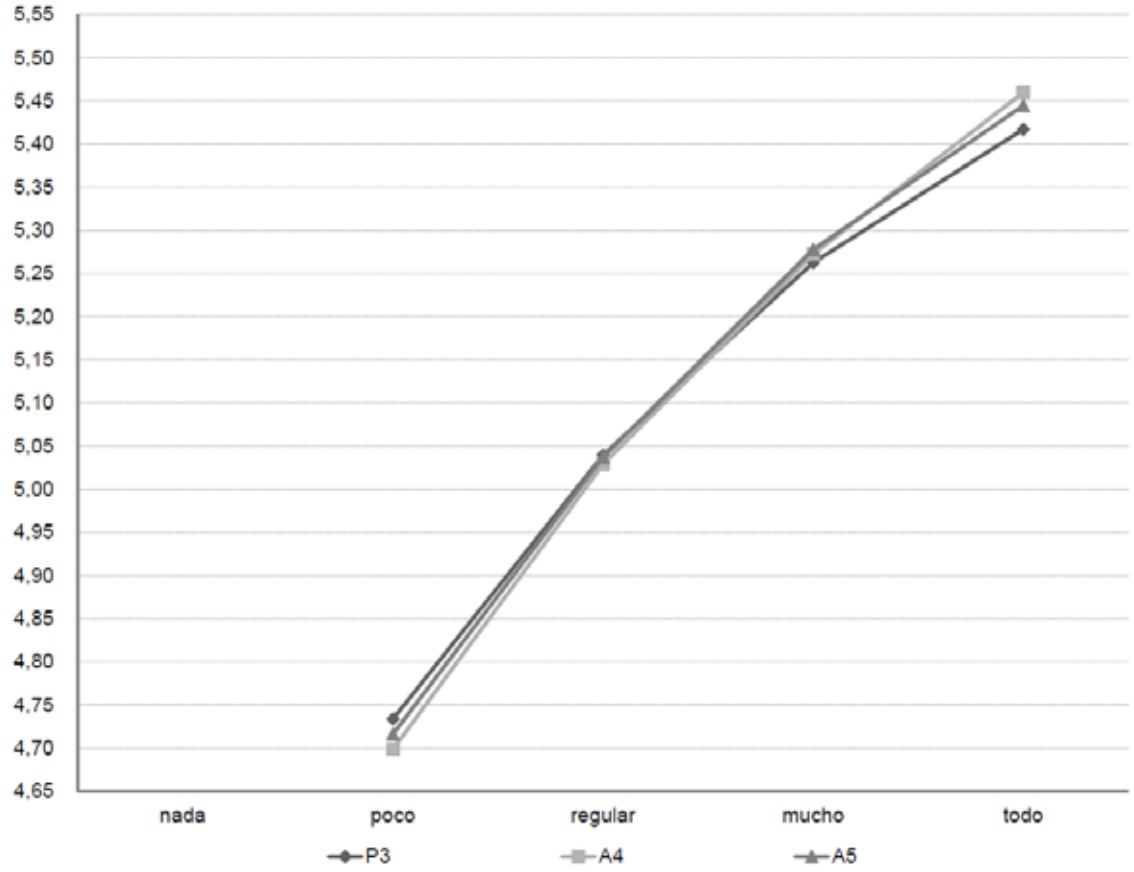


Fig. 4.9. Representación de la Viabilidad Global al cambiar el valor de las características de la categoría Problema de Negocio.

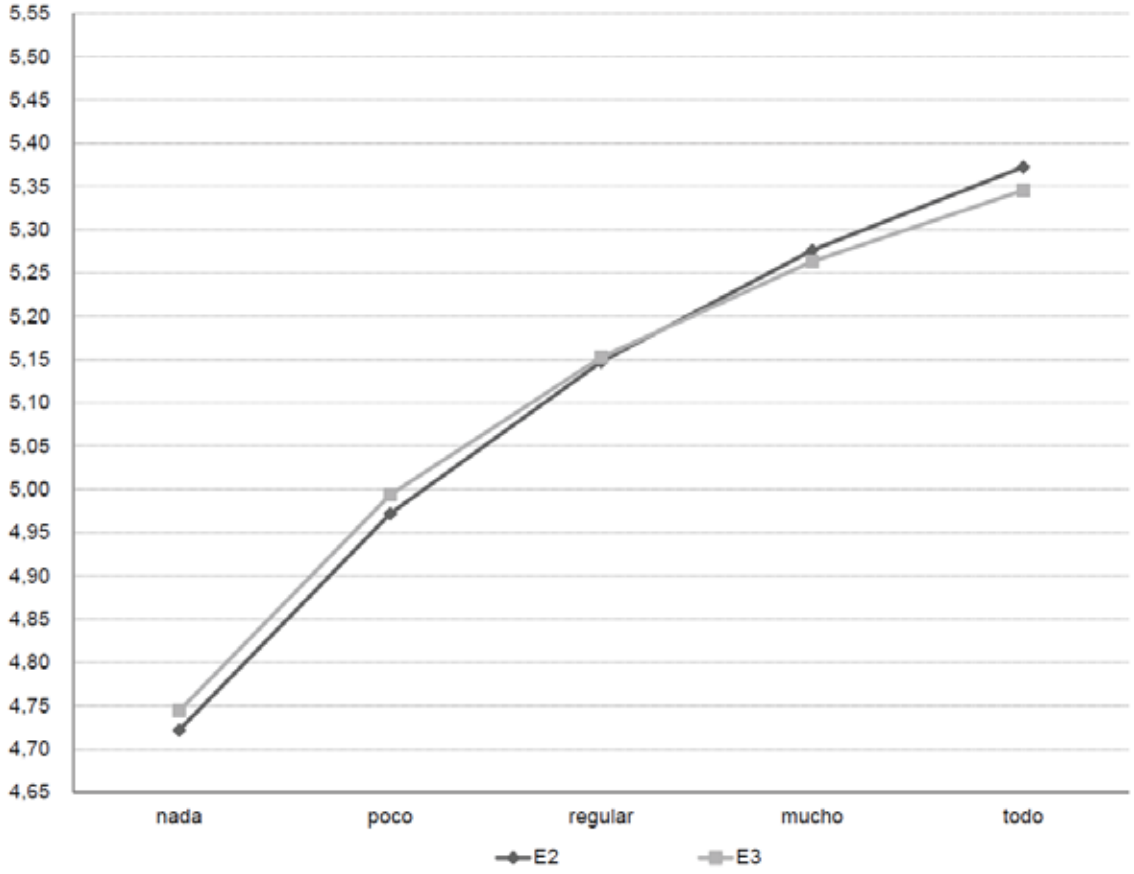


Fig. 4.10. Representación de la Viabilidad Global al cambiar el valor de las características de la categoría Tipo del Proyecto.

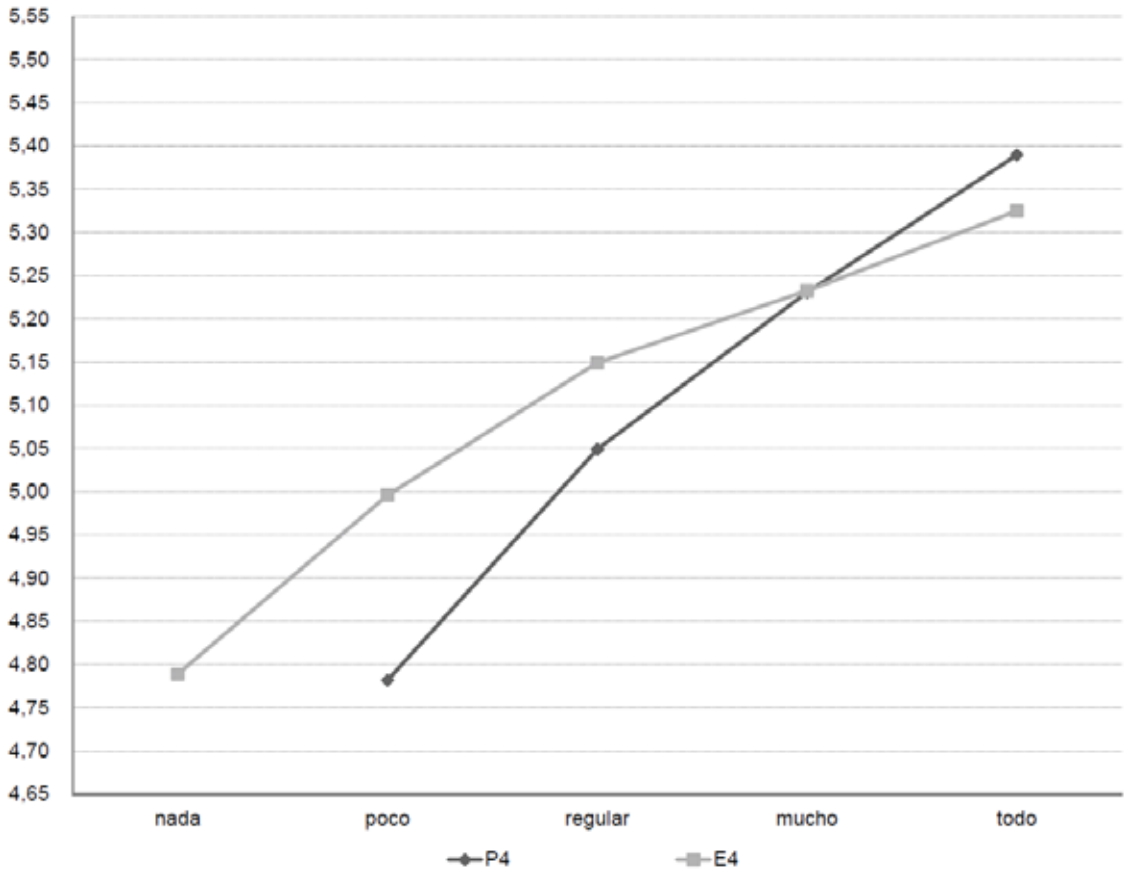


Fig. 4.11. Representación de la Viabilidad Global al cambiar el valor de las características de la categoría Equipo de Trabajo.

4.2.3.2.3. Análisis Estadístico por Viabilidad

Para finalizar este análisis estadístico, se generan dos gráficos donde se distribuye la cantidad de valores por característica teniendo en cuenta si el proyecto es viable o no (Figura 4.12). Debido a que varias características deben superar un umbral mínimo (igual a '*poco*'), se ha considerado en el gráfico a los valores '*nada*' y '*poco*' como uno sólo.

De lo observado en la Figura 4.12 (la cual está formado por los gráficos 4.12a y la 4.12b), se confirma lo indicado anteriormente para las dimensiones: a mayor valor de las características, la probabilidad de que el proyecto sea viable es mayor.

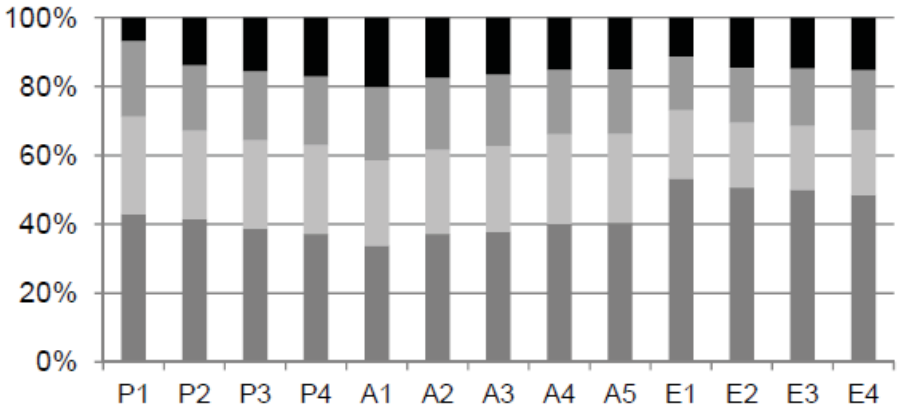
Por otra parte, se pueden inferir las siguientes conclusiones:

- del gráfico 4.12a donde se muestra la distribución de *proyectos no viables*, se resalta la característica que mide el grado de actualización de los datos disponibles (P1) por tener la menor cantidad de proyectos no viables cuando el valor tiende a '*todo*'.
- del gráfico 4.12b para la distribución de *proyectos viables* se resaltan tres características que producen que el proyecto sea viable a pesar de tener bajos valores. Éstas son el grado de representatividad de los datos (P2), el grado de aplicación de técnicas estadísticas

tradicionales para resolver el problema de negocio (A4) y el grado de estabilidad del problema de negocio (A5).

Asimismo al realizar un análisis en términos comparativos de ambos gráficos, cabe notar que la variación de las características asociadas al éxito no genera una diferencia significativa en la viabilidad del proyecto; no obstante en caso de que el valor de las características tiende a '*nada*', dicho proyecto tiende a no ser viable.

4.12a ***-*** *Distribución para Proyectos No Viables*



4.12b ***-*** *Distribución para Proyectos Viables*

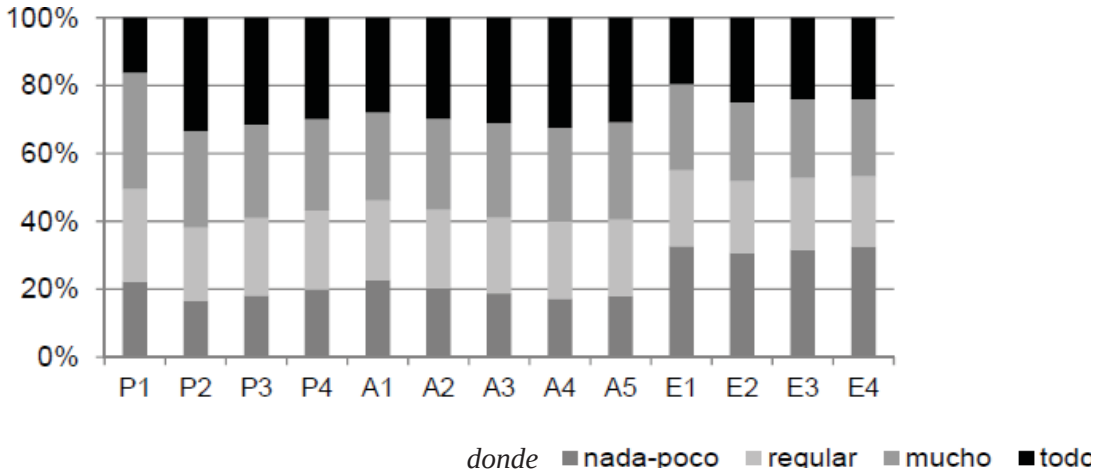


Fig. 4.12. Distribución de la cantidad de valores de las características de acuerdo a si el proyecto es viable o no.

4.2.3.2.4. Conclusiones del Análisis Estadístico para el Modelo de Viabilidad

A partir del estudio estadístico realizado se puede concluir que a mayor valor de las características (es decir, cuando el valor de éstas tiende a '*todo*'), crece las posibilidades de que el proyecto sea viable. En caso de que este hecho no sea posible, el ingeniero de Explotación de Información debe procurar aumentar el valor de las características asociadas a los *datos* y el *problema de negocio*. Dentro de estas categorías se destacan las siguientes características: el grado de representatividad de

los datos, el grado de aplicación de técnicas estadísticas tradicionales para resolver el problema de negocio, el grado de estabilidad del problema de negocio y si la tecnología de los repositorios facilita la manipulación de los datos. En otros términos, el ingeniero debería intentar obtener una mayor valoración de ellas para así asegurar la viabilidad del proyecto.

4.3. MODELO PARA LA ESTIMACIÓN DE ESFUERZO DE PROYECTOS EXPLOTACIÓN DE INFORMACIÓN

En esta sección se presenta la propuesta del modelo que permite estimar el esfuerzo para llevar a cabo en forma completa un proyecto de Explotación de Información dentro de una PyME, la cual se estructura en tres partes: generalidades del modelo (sección 4.3.1), propuesta del modelo (sección 4.3.2) y análisis preliminar del mismo (sección 4.3.3).

4.3.1. Generalidades del Modelo para la Estimación de Esfuerzo

De acuerdo al segundo problema identificado en el capítulo 3 de este trabajo de tesis, se considerar hacer mención nuevamente a la necesidad detectada sobre contar con un método de estimación que sea fiable para proyectos de Explotación de Información de tamaño pequeño o mediano, usualmente requeridos por las PyMEs [García-Martínez *et al.*, 2011b]. A pesar de la existencia de un modelo para estimar el esfuerzo en proyectos de Explotación de Información propuesto en [Marbán *et al.*, 2008], el cual se denomina DMCoMo; conforme a lo demostrado en [Pytel *et al.*, 2011], dicho modelo sólo es confiable para proyectos grandes. En consecuencia, cuando es aplicado en proyectos pequeños o de medianos porte, se ha comprobado que genera sobrestimaciones muy grandes. Las estimaciones mínimas obtenidas para DMCoMo en el estudio antes mencionado se encuentran en el orden de los 33 meses/hombre (es decir, casi 3 años/hombre), mientras que un proyecto pequeño o mediano de Explotación de Información se suele culminar en menos de 18 meses/hombre.

El modelo propuesto en este trabajo de tesis se basa en los métodos de estimación para proyectos de Ingeniería de Software de la familia COCOMO [Boehm *et al.*, 2000]. En este sentido, incluye la definición de un conjunto de factores de costo que permiten caracterizar al proyecto. Dichos factores de costo se han agrupados de acuerdo a qué o a quién se refiere; siendo importante resaltar, que los mismos son similares, pero no idénticos, a las condiciones aplicadas en el modelo para la evaluación de la viabilidad descripto en la Sección 4.2.2. Esto permite al Ingeniero de Explotación de Información capitalizar nuevamente gran parte de la información que ya ha sido recolectada sobre el proyecto. En la metodología CRISP-DM, el modelo propuesto debe ser utilizado durante la actividad '*Costos y beneficios*' correspondiente a la tarea general '*Evaluar la Situación*' de la fase

'*Comprensión del Negocio*'. La ubicación de dicha tarea específica se puede visualizar en la Figura 4.13. En cambio, si se utiliza el Modelo de Proceso, esto se debe llevar a cabo en la tarea '*Calcular el costo estimado del proyecto*' del subproceso '*Planificación / Entendimiento del Negocio*' dentro del proceso general de '*Administración de Proyectos*' (Figura 4.14).

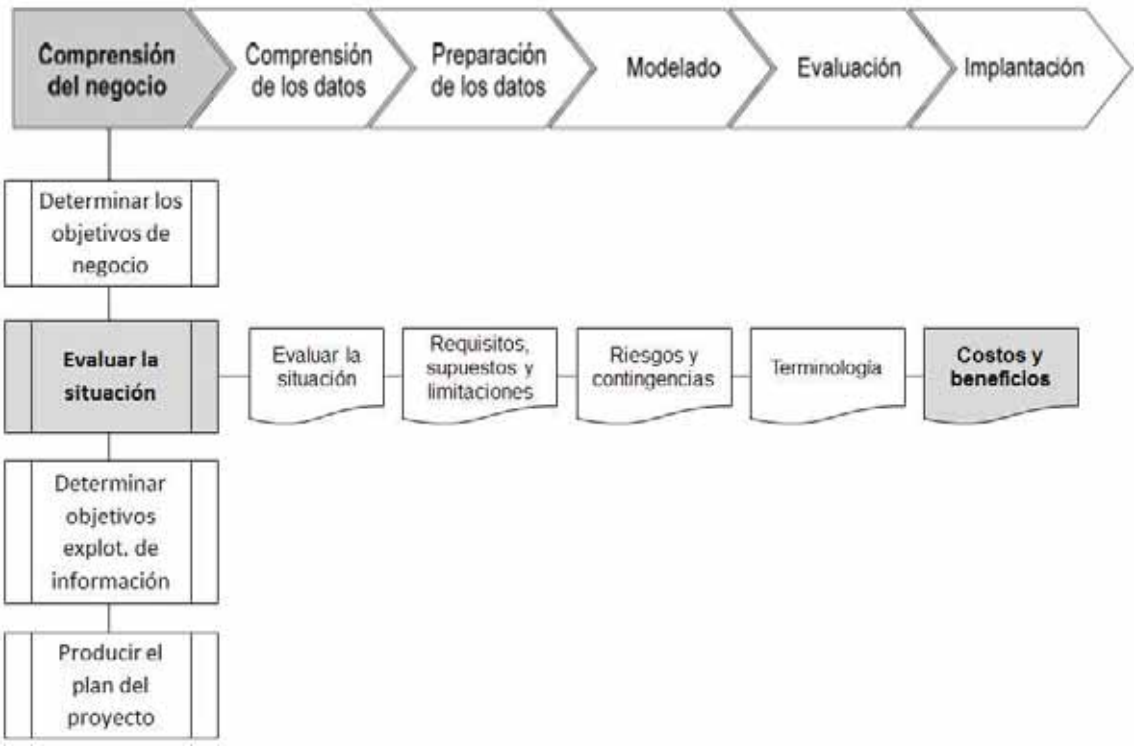


Fig. 4.13. Ubicación de la tarea de la metodología CRISP-DM en la que se aplica el Modelo de Estimación propuesto.

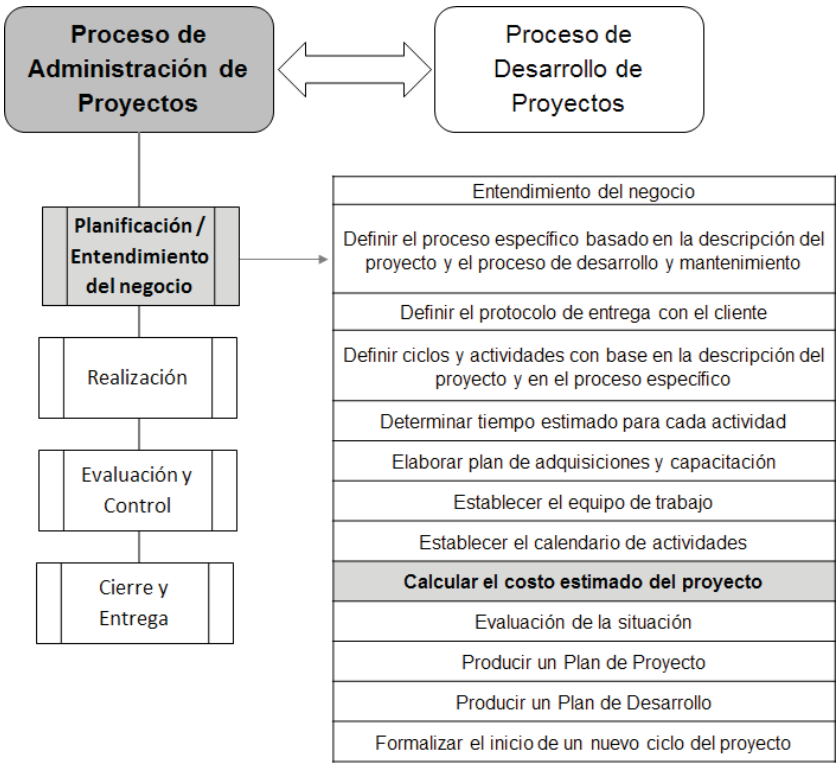


Fig. 4.14. Ubicación de la tarea del Modelo de Proceso en la que se aplica el Modelo de Estimación propuesto.

4.3.2. Propuesta del Modelo para la Estimación de Esfuerzo

La propuesta del Modelo de Estimación de Esfuerzo para PyMEs incluye la identificación de los factores de costo asociados a las características del proyecto (sección 4.3.2.1). Estos factores de costo son utilizados por los dos métodos especificados para realizar la estimación de esfuerzo: uno que aplica una fórmula lineal obtenida mediante regresión (sección 4.3.2.2) y otro, más complejo, que aplica la combinación de los factores de costo con una fórmula obtenida de forma empírica (sección 4.3.2.3).

En forma similar a como se ha explicado el modelo anterior, en este caso también se propone un proceso que permite realizar la estimación de esfuerzo, el cual se compone de sólo tres pasos que son los siguientes:

1. Determinar el valor correspondiente para cada uno de los factores de costo del proyecto.
2. Utilizar los valores de los factores de costo en el método seleccionado.
3. Calcular el esfuerzo estimado aplicando la fórmula correspondiente.

Como consecuencia de la aplicación de este proceso, se obtiene la cantidad de meses/hombre que son precisos para llevar a cabo el proyecto de manera completa.

4.3.2.1. Factores de Costo para la Estimación de Esfuerzo

Teniendo en cuenta las características de los proyectos de Explotación de Información para PyMEs que se indican en capítulo 2 de este trabajo de tesis, se definen ocho factores de costos para ser evaluados por el modelo. Se han definido pocos factores de costo, debido a que como se demuestra en [Chen *et al.*, 2005] al momento de crear un nuevo método de estimación es preferible ignorar muchos de los datos no significativos para evitar así que sea demasiado complejo y, por consiguiente, poco práctico. De este modo, se busca eliminar tanto las variables irrelevantes como las que depende de otras y, a su vez, reducir la varianza y el ruido en los resultados.

Asimismo, los factores de costo han sido seleccionados teniendo en cuanta las tareas más críticas de la metodología CRISP-DM:

- En [Yang *et al.*, 2006] se indica que actualmente la construcción de los modelos de Minería de Datos y búsqueda de patrones es bastante simple, pero el 90% de los esfuerzos del proyecto están incluidos en el pre-procesamiento de los datos (es decir, la fase de '*Preparación de los Datos*' de CRISP-DM).

- A partir de nuestra experiencia [Rodríguez *et al.*, 2010], las otras tareas críticas se relacionan con la fase de '*Comprensión del Negocio*' (entre las que se destacan el entendimiento del negocio y la identificación de las metas del proyecto).

Los factores de costos propuestos se encuentran agrupados en tres grupos dependiendo de su naturaleza como se indica a continuación:

- **Factores de costo relacionados al Tipo de Proyecto:**

  - Tipo de objetivo de explotación de información (OBTY)

    Este factor de costo analiza el objetivo del proyecto a partir del tipo de proceso de Explotación de Información a ser aplicado. Los valores de este factor de costo han sido definidos de acuerdo a la definición de los procesos realizada en [Britos & García-Martínez, 2009; García-Martínez *et al.*, 2011a] como se puede observar en la Tabla 4.7.

Valor	Descripción
1	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida.
2	Se desea dividir los datos disponibles en grupos sin poseer una clasificación conocida previamente.
3	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente.
4	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida.
5	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre la identificación de una clase desconocida previamente.

Tabla 4.7. Valores del factor de costo OBTY.

  - Grado de apoyo de los miembros de la organización (LECO)

    El grado de apoyo y participación de los miembros de la organización se analiza por cada nivel de la organización. Se debe considerar en qué medida la gerencia media (supervisores y jefes de área) y/o el resto del personal están dispuestos a asistir al equipo de trabajo en el entendimiento del negocio y los datos disponibles.

    Se sobreentiende que si un proyecto se encuentra en la etapa de planificación se va a contar con el apoyo de la alta gerencia (normalmente los dueños de la PyME).

    Los valores definidos para este factor de costo se indican en la Tabla 4.8.

Valor	Descripción
1	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto.
2	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto.
3	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente.
4	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar.

Tabla 4.8. Valores del factor de costo LECO.

- **Factores de costo relacionados a los Datos Disponibles:**

  - Cantidad y tipo de los repositorios de datos disponibles (AREP)

    Aquí se analizan las fuentes de datos disponibles (es decir, sistemas gestores de bases de datos, planillas de cálculos, documentos, entre otros). En este caso, interesa saber tanto la cantidad de repositorios (públicos o privados de la organización) como la tecnología en que se encuentran implementadas. No es de interés conocer la cantidad de tablas que posee cada repositorio, debido a que la integración de las mismas dentro de un repositorio es relativamente sencilla. En particular, cuando se hace uso de sistemas gestores de bases de datos que permiten integrar las tablas con un simple comando query.

    Sin embargo, en función de la tecnología empleada, la complejidad de las tareas de integración entre repositorios puede ser mayor o menor. Los criterios recomendados para identificar la compatibilidad de integración entre repositorios son los siguientes:

    - Si todos los repositorios están implementados con la misma tecnología, entonces se consideran como compatibles para la integración.
    - Si todos los repositorios permiten exportar los datos en un formato común, entonces pueden ser considerados como compatibles para la integración (en este caso se realizaría la integración con los datos exportados).
    - Por otro lado, si existen repositorios que no están en forma digital (es decir impreso en papel) se considera que la tecnología es no compatible. En esta situación se debe tener en cuenta que el método de estimación no puede predecir el tiempo requerido para realizar la digitalización de esta información, ya que esto puede variar de acuerdo a muchos factores externos (como son la longitud, diversidad, formato entre otros).

    La tabla con los valores de este factor de costo se indica en la Tabla 4.9.

Valor	Descripción
1	Sólo 1 repositorio disponible.
2	Entre 2 y 4 repositorios con tecnología compatible para la integración.
3	Entre 2 y 4 repositorios con tecnología no compatible para la integración.
4	Más de 5 repositorios con tecnología compatible para la integración.
5	Más de 5 repositorios con tecnología no compatible para la integración.

Tabla 4.9. Valores del factor de costo AREP.

- Cantidad de tuplas disponibles en la tabla principal (QTUM)

  Este factor de costo evalúa la cantidad total de tuplas (registros) disponibles en la tabla principal a ser utilizada en la aplicación del proceso de explotación de información.

  Los posibles valores de este factor de costo se indican en la Tabla 4.10.

Valor	Descripción
1	Hasta 100 tuplas en la tabla principal.
2	Entre 101 y 1.000 tuplas en la tabla principal.
3	Entre 1.001 y 20.000 tuplas en la tabla principal.
4	Entre 20.001 y 80.000 tuplas en la tabla principal.
5	Entre 80.001 y 5.000.000 tuplas en la tabla principal.
6	Más de 5.000.000 tuplas en la tabla principal.

Tabla 4.10. Valores del factor de costo QTUM.

- Cantidad de tuplas disponibles en las tablas auxiliares (QTUA)

  Esta variable considera la cantidad aproximada de tuplas (registros) disponibles en las tablas auxiliares utilizadas para agregar información complementaria a la tabla principal. Ejemplos de tablas auxiliares son la tabla que describen las características de los productos vendidos o la tabla que almacena en forma centralizada los datos de cada cliente. Estas tablas (si existen y se utilizan) normalmente suelen tener menos registros que la tabla principal, lo cual es considerado por los valores del factor de costo en la Tabla 4.11.

Valor	Descripción
1	No se utilizan tablas auxiliares.
2	Hasta 1.000 tuplas en las Tablas auxiliares.
3	Entre 1.001 y 50.000 tuplas en las Tablas auxiliares.
4	Más de 50.000 tuplas en las Tablas auxiliares.

Tabla 4.11. Valores del factor de costo QTUA.

- Nivel de conocimiento sobre los datos (KLDS)

  Otro factor a tener presente es el nivel de documentación existente sobre los repositorios de datos. Se debe analizar si existe un documento donde se explique la tecnología en que los repositorios están implementados, los atributos que componen sus tablas y la forma en que los datos son creados, modificados, y/o eliminados. En caso de que esta información no se encuentre disponible, es necesario realizar mayor cantidad de reuniones con los encargados de la administración y mantenimiento de los repositorios para que sean aclaradas.

  En la Tabla 4.12 se indican los valores para este factor de costo.

Valor	Descripción
1	Todas las tablas y repositorios están correctamente documentados.
2	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos.
3	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos.
4	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos.
5	Las tablas y repositorios no están documentados y existen expertos en los datos pero no están disponibles para explicarlos.
6	Las tablas y repositorios no están documentados y no existen expertos en los datos para explicarlos.

Tabla 4.12. Valores del factor de costo KLDS.

- **Factores de costo relacionados a los Recursos Disponibles:**

  - Nivel de conocimiento y experiencia del equipo de trabajo (KEXT)

    Analiza la capacidad del equipo de trabajo que se ocupa de llevar adelante el proyecto. Obviamente, el equipo de trabajo contratado para realizar el proyecto debe tener un mínimo conocimiento y experiencia en el desarrollo de proyectos de Explotación de Información. No obstante, pueden poseer o no experiencia en proyectos con objetivos similares, dentro del mismo tipo de negocio y/o usando datos similares a los actuales.

    Por consiguiente, estos tres elementos son evaluados por el factor de costo (Tabla 4.13) ya que afectan el esfuerzo requerido para completar el proyecto.

  - Funcionalidad de las herramientas disponibles (TOOL)

    Finamente, este factor de costo evalúa las características de las herramientas disponibles para ser aplicadas en el proyecto. A tal efecto, se analizan las funcionalidades que poseen tanto para la preparación (limpieza, integración y formateo) como el modelado de los datos

(los algoritmos de minería de datos implementados). Sus posibles valores se indican en la Tabla 4.14.

Valor	Descripción
1	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos.
2	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos.
3	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos.
4	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos.
5	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos.

Tabla 4.13. Valores del factor de costo KEXT.

Valor	Descripción
1	La herramienta posee funciones tanto para el formateo e integración de los datos (permitiendo importar más de una Tabla de datos) como para aplicar las técnicas de minería de datos.
2	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, y permite importar más de una Tabla de datos en forma independiente.
3	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una Tabla de datos.
4	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos, y permite importar más de una Tabla de datos.
5	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una Tabla de datos.

Tabla 4.14. Valores del factor de costo TOOL.

4.3.2.2. Especificación de la Fórmula Lineal para la Estimación de Esfuerzo

Una vez que los factores de costo han sido definidos, se han utilizado para caracterizar treinta y cuatro proyectos de Explotación de Información provistos por colegas investigadores, junto con su esfuerzo real asociado. Con los datos de estos proyectos (que se incluyen en el Anexo A de este trabajo), se ha aplicado un método de regresión lineal multivariante [Weisberg, 1985] para obtener una ecuación lineal. Esta fórmula permite calcular el esfuerzo en meses/hombre para desarrollar el proyecto. En otros términos, permite calcular el tiempo en meses que le llevaría a una sola persona realizar el proyecto en forma completa.

La fórmula lineal propuesta es la siguiente:

$$PEM_L = 0{,}80 \cdot OBTY + 1{,}10 \cdot LECO - 1{,}20 \cdot AREP - 0{,}30 \cdot QTUM - 0{,}70 \cdot QTUA + 1{,}80 \cdot KLDS - 0{,}90 \cdot KEXT + 1{,}86 \cdot TOOL - 3{,}30$$

Donde:

PEM_L: es el esfuerzo en meses/hombre estimado por la fórmula lineal de estimación.

OBTY, *LECO*, *AREP*, *QTUM*, *QTUA*, *KLDS*, *KEXT* y *TOOL*: son los valores correspondientes de los factores de costo definidos en las Tablas 4.7 a 4.14 respectivamente.

4.3.2.3. Especificación del Método Empírico para la Estimación de Esfuerzo

Debido a los problemas detectados (los cuales se explican en las secciones 4.3.3.2.1 y 4.3.3.2.2) en el comportamiento general de la fórmula lineal especificada en la sección anterior, se ha propuesto un segundo método de estimación para este modelo. Este método es similar a los métodos de la familia COCOMO disponibles para proyectos de la Ingeniería de Software.

Teniendo en cuenta los mismos factores de costo que se detallaron anteriormente en la sección 4.3.2.1, se ha solicitado la opinión de investigadores expertos en el dominio y se han llevado a cabo simulaciones mediante el método Monte Carlo (cuyos resultados finales se describen en la sección 4.3.3.2.3); para determinar de qué manera, la combinación de dichos factores puede afectar el esfuerzo requerido. A partir de la combinación de los factores de costo resultante es posible determinar tres coeficientes generales para caracterizar al proyecto: *complejidad del negocio*, *complejidad de los datos* y *complejidad de las técnicas de modelado*. A su vez, se define un *coeficiente de ajuste* que depende del nivel de experiencia y conocimientos del equipo de trabajo. Todos estos valores son luego utilizados en una nueva fórmula que permite calcular el esfuerzo requerido en meses/hombre.

De esta forma, para estimar el esfuerzo aplicando el método empírico, en el paso 2 del proceso indicado en la sección 4.3.2 se deben realizar las siguientes acciones:

a) Determinar el valor del coeficiente asociado a la complejidad del negocio.

En primer término, es necesario establecer el coeficiente de complejidad asociado a las tareas de relevamiento y análisis de la organización donde se desarrolla el proyecto (fase '*Comprensión del Negocio*' de la metodología CRISP-DM). A tal efecto, se consideran los factores de costo relacionados al tipo de objetivo del proyecto (OBTY), el grado de apoyo

de los miembros de la organización (LECO), y el nivel de conocimiento disponible sobre los datos (KLDS).

Estos tres factores de costo se combinan en la tabla de decisión 4.15 para obtener el coeficiente de complejidad del negocio (CNEG) correspondiente.

Valores LECO	**Valores OBTY**	**Valores KLDS**	
		< 3	**≥ 3**
= 1	**-**	*1,00*	*2,00*
≥ 2	**= 1**	*2,00*	*3,70*
	≥ 2	*3,70*	

Tabla 4.15. Tabla de Decisión para determinar CNEG.

b) Determinar el valor del coeficiente asociado a la complejidad de los datos.

En una segunda instancia, se identifica la complejidad asociada al entendimiento y la acción de transformar los datos en el desarrollo del proyecto (correspondientes a las fases de '*Comprensión de los Datos*' y '*Preparación de los Datos*'). Esto se logra considerando los factores de costo correspondientes a la cantidad de tuplas disponibles en la tabla principal (QTUM), la cantidad de tuplas disponibles en las tablas auxiliares (QTUA), y las características de los repositorios disponibles (AREP). Estos factores de costo se combinan en la tabla de decisión 4.16 para determinar la complejidad de los datos (CDAT).

Valores QTUM	**Valores QTUA**	**Valores AREP**	
		= 1	**≥ 2**
= 1	**-**	*0,25*	
= 2	**-**	*0,50*	
= 3	**= 1**	*1,50*	*2,40*
	≥ 2	*2,40*	
≥ 4	**-**	*2,40*	

Tabla 4.16. Tabla de Decisión para determinar CDAT.

c) Determinar el valor del coeficiente asociado a la complejidad del modelado.

En esta etapa se evalúa la complejidad de las técnicas de modelado a ser aplicadas sobre los datos para obtener los resultados del proyecto. Este coeficiente afecta a las tres últimas fases

de la metodología CRISP-DM. Esto significa que se debe considerar nuevamente el tipo de objetivo del proyecto (OBTY) junto con las funcionalidades provistas por la herramienta disponible (TOOL). Estos factores se combinan en la tabla de decisión 4.17 obteniendo así el coeficiente de modelado (CMOD).

Valores TOOL	**Valores OBTY**	
	= 1	**≥ 2**
< 4	*0,60*	*0,80*
≥ 4	*2,70*	*3,80*

Tabla 4.17. Tabla de Decisión para determinar CMOD.

d) Determinar el valor del coeficiente de ajuste por la experiencia del equipo de trabajo.

Finalmente, para establecer el coeficiente de ajuste por la experiencia y conocimientos del equipo de trabajo (AEXP), sólo es necesario analizar el factor de costo KEXT tal como se indica en la Tabla 4.18.

Valores KEXT	**Coeficiente de Ajuste por Experiencia (AEXP)**
= 1	*0,60*
= 2	*0,70*
= 3	
= 4	*0,80*
= 5	*1,00*

Tabla 4.18. Tabla de Decisión para determinar AEXP.

A partir de los valores obtenidos en los pasos anteriores para los coeficientes CNEG, CDAT, CMOD y AEXP, es posible aplicar una fórmula obtenida empíricamente para calcular el esfuerzo en meses/hombre. La fórmula empírica de estimación es la siguiente:

$$PEM_E = (1{,}80 \cdot CNEG + 0{,}90 \cdot CDAT + 1{,}40 \cdot CMOD - 1{,}50) \cdot AEXP$$

Donde:

PEM_E: es el esfuerzo en meses/hombre estimado por la fórmula empírica de estimación.

CNEG, *CDAT*, *CMOD* y *AEXP*: son los valores correspondientes obtenidos en los pasos anteriores.

4.3.3. Análisis Preliminar del Modelo para la Estimación de Esfuerzo

En esta sección se presenta un análisis preliminar del modelo de estimación propuesto (téngase en cuenta que un análisis más completo para su validación se describe en el capítulo siguiente). Este análisis preliminar busca ilustrar la utilización del modelo en un caso de prueba (sección 4.3.3.1), así como, analizar estadísticamente el comportamiento de cada método a través de su aplicación en proyectos simulados (sección 4.3.3.2).

4.3.3.1. Prueba de Concepto del Modelo para la Estimación de Esfuerzo

El caso de prueba utilizado para el modelo para estimación de esfuerzo es el mismo que fue aplicado en la prueba de concepto positiva del modelo para evaluación de la viabilidad (sección 4.2.3.1). Debido a que dicho proyecto había finalizado con éxito, es posible comparar el esfuerzo real que fue requerido para llevarlo a cabo de forma completa con el esfuerzo estimado por cada método del modelo propuesto. El proyecto fue desarrollado por 3 personas en 4 meses, por lo que es posible indicar que tuvo un esfuerzo real de 12 meses/hombre (o, lo que es lo mismo, un año/hombre). Asimismo, a partir de las características del proyecto, se han definidos los valores de los factores de costo del modelo (descriptos en la sección 4.3.2.1) según se indica en la Tabla 4.19.

Categoría	ID	Descripción	Valor
Tipo del Proyecto	OBTY	*Se desea conocer la incidencia de los atributos sobre el motivo de baja del servicio.*	5
	LECO	*Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto.*	2
Datos Disponibles	AREP	*Se disponen entre 2 y 4 repositorios con tecnología no compatible para la integración*	3
	QTUM	*Hay aproximadamente 65.000 tuplas en la Tabla principal*	4
	QTUA	*Hay aproximadamente 10.000 tuplas en Tablas auxiliares*	3
	KLDS	*Las Tablas y repositorios no están documentados y no existen expertos disponibles.*	6
Recursos Disponibles	KEXT	*El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos.*	4
	TOOL	*La herramienta utilizada (TANAGRA) posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una Tabla de datos.*	5

Tabla 4.19. Datos del proyecto para la prueba de concepto.

Con estos datos, se procede a realizar la estimación del esfuerzo aplicando la fó6rmula lineal:

$$PEM_L = 0{,}80 \cdot OBTY + 1{,}10 \cdot LECO - 1{,}20 \cdot AREP - 0{,}30 \cdot QTUM - 0{,}70 \cdot QTUA + 1{,}80 \cdot KLDS - 0{,}90 \cdot KEXT + 1{,}86 \cdot TOOL - 3{,}30$$

$$PEM_L = (0{,}80 \cdot 5) + (1{,}10 \cdot 2) - (1{,}20 \cdot 3) - (0{,}30 \cdot 4) - (0{,}70 \cdot 3) + (1{,}80 \cdot 6) - (0{,}90 \cdot 4) + (1{,}86 \cdot 5) - 3{,}30$$

$$PEM_L = 12{,}50 \text{ meses/hombre}$$

Como se puede ver, el resultado de aplicar la fórmula lineal obtiene un esfuerzo estimado de 12,50 meses/hombre. Si se compara con el esfuerzo real del proyecto (12 meses/hombre), se observa que existe un error muy pequeño de sólo medio meses/hombre (es decir, aproximadamente 15 días/hombre). Esto significa que para este ejemplo el método lineal fue muy preciso.

De manera alternativa, usando los mismos factores de costo de la Tabla 4.19, se aplica el método empírico. En primer término, se determinan los valores de los coeficientes correspondientes como se muestra en la Tabla 4.20.

Coeficiente	**Factor de Costo**		**Valor Asignado al Coeficiente**
	ID	**Valor**	
Complejidad del Negocio (CNEG)	OBTY	5	*3,70*
	LECO	2	
	KLDS	6	
Complejidad de los Datos (CDAT)	QTUM	4	*2,40*
	QTUA	3	
	AREP	3	
Complejidad del Modelado (CMOD)	OBTY	5	*3,80*
	TOOL	5	
Ajuste por Experiencia del Equipo (AEXP)	KEXT	4	*0,80*

Tabla 4.20. Valores de los coeficientes del método empírico para la prueba de concepto.

Estos coeficientes se aplican en la fórmula correspondiente como se indica a continuación:

$$PEM_E = \left(1{,}80 \cdot CNEG + 0{,}90 \cdot CDAT + 1{,}40 \cdot CMOD - 1{,}50\right) \cdot AEXP$$

$$PEM_E = \left[(1{,}80 \cdot 3{,}70) + (0{,}90 \cdot 2{,}40) + (1{,}40 \cdot 3{,}80) - 1{,}50 \right] \cdot 0{,}80$$

$$PEM_E = 12{,}64 \cdot 0{,}80$$

$$PEM_E = 10{,}11 \text{ meses/hombre}$$

El resultado obtenido es un poco menos preciso en este caso por estimar un esfuerzo de 10,11 meses/hombre. El error generado es de 1,89 meses/hombre (aproximadamente 57 días/hombre). A pesar de esta diferencia, el método se considera confiable en esta instancia para continuar con su análisis estadístico en la siguiente sección.

4.3.3.2. Análisis Estadístico del Modelo para la Estimación de Esfuerzo

Para analizar el comportamiento de los métodos de estimación propuestos en este modelo, se ha usado nuevamente el método de simulación Monte Carlo [Metropolis & Ulam, 1949; Kalos & Withlock, 1986]. En forma pseudo-aleatoria se han generado combinaciones de los valores para los ocho factores de costo obteniendo un banco de pruebas con 40.000 proyectos que se encuentran disponibles en [Pytel, 2013a]. Con estos datos se han aplicado los dos métodos provistos por el modelo para calcular las estimaciones correspondientes; es decir, se utiliza tanto la fórmula de estimación lineal como el método de estimación empírico sobre dichos datos. En primer término, las estimaciones obtenidas se analizan de manera general comparando los resultados de cada método (sección 4.3.3.2.1); y posteriormente se lleva a cabo un estudio particular y en detalle, tanto para la fórmula de estimación lineal (sección 4.3.3.2.2) como para el método de estimación empírico (sección 4.3.3.2.3). Finalmente, teniendo en cuenta los resultados obtenidos se presentan las conclusiones preliminares del comportamiento de cada método (sección 4.3.3.2.4).

4.3.3.2.1. Análisis Estadístico General

En primer término se lleva a cabo un análisis general estadístico estudiando la distribución del esfuerzo estimado por cada método del modelo. Los resultados estadísticos obtenidos se indican en la Tabla 4.21.

Parámetro Estadístico	Método de Estimación	
	Fórmula Lineal (PEM_L)	**Método Empírico (PEM_E)**
Valor Mínimo	*-12,54*	*0,82*
Valor Máximo	*22,10*	*12,64*
Valor Promedio	*4,64*	*6,29*
Valor de Varianza	*25,98*	*5,63*

Tabla 4.21. Resultados estadísticos (en meses/hombre) para cada método de estimación.

Para facilitar la interpretación de estos resultados estadísticos, se representan en el gráfico Boxplot [Turkey, 1977] de la Figura 4.15. Este tipo de gráfico permite representar en una única figura los datos correspondientes a los límites superior e inferior (valores máximo y mínimo), el desvío máximo (media más la desviación estándar) y mínimo (media menos la desviación estándar), así como la media (o promedio) de las estimaciones obtenidas.

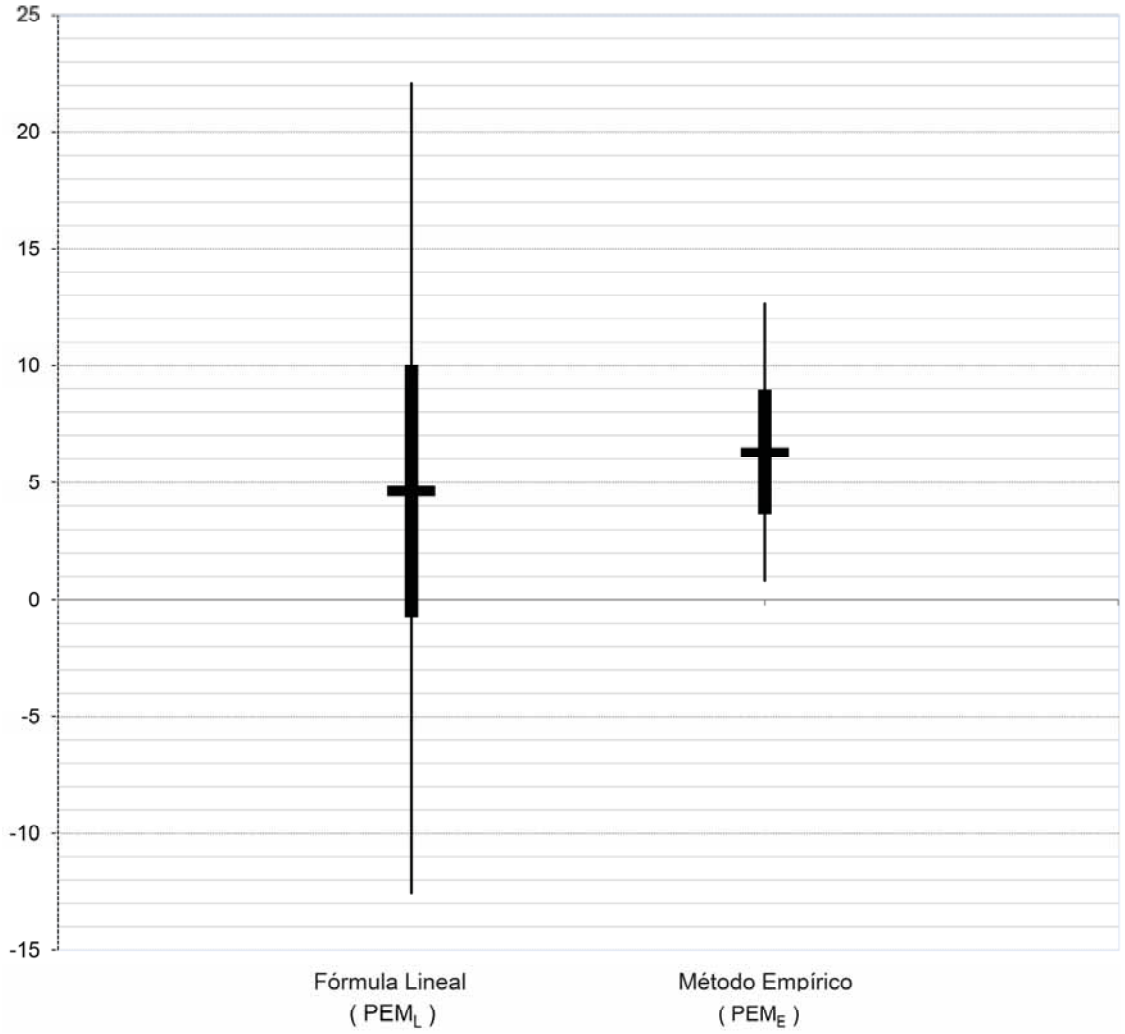


Fig. 4.15. Gráfico boxplot comparando el comportamiento general de los métodos de estimación propuestos.

Al realizar la primera observación de los resultados generales obtenidos, se nota una falla en las estimaciones realizadas por la fórmula lineal. Este método ha generado esfuerzos menores o iguales a cero, lo cual no proporciona resultados coherentes con la realidad. Esto significa que para ciertos proyectos la fórmula lineal presenta ciertas falencias, al considerar que la realización del proyecto no insume ningún esfuerzo para el equipo de desarrollo. De todas formas, se trata de una pequeña cantidad de casos debido a que sólo sucede en el 19% de los proyectos generados (exactamente 7.631 de los 40.000 proyectos generados). Este inconveniente tiene lugar en función de la combinación de ciertos valores de los factores de costo, los cuales se indican a continuación:

- valores menores a 4 para el factor de costo asociado al nivel de conocimiento sobre los datos (KLDS), que significa que en la organización existe documentación disponible sobre las fuentes de datos.
- valores mayores a 2 para el factor de costo sobre la cantidad y tipo de los repositorios de datos disponibles (AREP), es decir, cuando los repositorios no son compatibles para la integración o hay más de 5 repositorios.
- valores mayores a 2 para el nivel de conocimiento y experiencia del equipo de trabajo del proyecto (KEXT), por lo que el equipo de trabajo no ha trabajado en organizaciones similares con anterioridad.
- valores menores a 4 para la funcionalidad de las herramientas disponibles (TOOL) que implica que la herramienta incluye funciones de formateo.

Luego de haber hablado con expertos en el campo de la Explotación de Información, se ha determinado, que en proyectos reales que tienen lugar en el campo de en una PyME, la probabilidad de encontrar esta combinación de valores es muy baja. Por consiguiente, se decide continuar el estudio de la fórmula lineal para aquellos proyectos que no presentan la combinación de valores de factores de costo que producen el inconveniente mencionado. Tal como se podrá ver en el capítulo 5 correspondiente a la Validación de la Solución propuesta en este trabajo de tesis, la fórmula lineal genera resultados satisfactorios para proyectos reales en el contexto de las PyMEs.

De todas formas, en caso de que se presente la necesidad de realizar la estimación de un proyecto que contenga la combinación de factores de costo señalada anteriormente, es posible llevar a cabo dicha estimación mediante la aplicación del método empírico que se propone en este trabajo de tesis. Como se puede observar en la Tabla 4.21, el mínimo esfuerzo producido por este último método es válido, de 0,82 meses/hombre (es decir, aproximadamente 25 días/hombre).

Continuando con esta línea de análisis, se observa que los valores de la fórmula lineal (PEM_L) se encuentran mucho más dispersos que los del método empírico (PEM_E). Mientras PEM_E presenta

un máximo de 12,64 meses/hombre, PEM_L tiene casi el doble con 22,10 meses/hombre. Este hecho pone de manifiesto que PEM_E es mucho más conservador que $PEM_{L,}$ en términos de que el primero proporciona estimaciones con un menor grado de dispersión. De todas formas, el rango comprendido por PEM_E se considera suficiente para proyectos cortos que son normalmente los requeridos por las PyMEs.

Para finalizar este primer análisis, se observa que el rango comprendido por el desvío mínimo y máximo presenta características similares, dado que el conjunto de valores de PEM_E (entre 3,92 y 8,66 meses/hombre respectivamente) está contenido por el conjunto correspondiente a PEM_L (entre -0,46 y 9,74 meses/hombre).

4.3.3.2.2. Análisis Estadístico de la Fórmula Lineal para la Estimación de Esfuerzo

Con el objetivo de estudiar con mayor detalle el comportamiento de la fórmula lineal, se analizan los gráficos que ilustran la modificación del esfuerzo estimado al variar el valor de cada uno de los factores de costo. Para facilitar su interpretación, los factores de costos se encuentran agrupados de acuerdo al elemento al que pertenecen (los cuales fueron presentados en la sección 4.3.2.1):

- En el primer gráfico de la Figura 4.16 se representa la variación de los factores de costo relacionados al *Tipo de Proyecto* donde se puede observar:
  - En el caso del factor de costo asociado al tipo de objetivo del proyecto (OBTY) se puede observar que el crecimiento del esfuerzo estimado es proporcional al cambio del valor del factor de costo (de aproximadamente 17% entre cada valor).
    Este resultado se puede considerar como correcto por tener este factor de costo asociado valores más pequeños para objetivos sencillos de conseguir e interpretar, y valores más grandes para objetivos más complejos, que por supuesto demandan un esfuerzo mayor.
  - Para el grado de apoyo de los miembros de la organización (LECO) sucede algo similar; cuanto menor es el apoyo suministrado por la organización (lo que se corresponde con valores mayores del factor de costo), el esfuerzo promedio de realizar el proyecto aumenta. Por otra parte, en el gráfico de la Figura 4.16, también se observa que mientras la alta dirección y la gerencia media poseen buena disposición al proyecto (que corresponde a los valores LECO=1 y LECO=2) el compromiso del personal de apoyo no genera una diferencia de esfuerzo significativa. En cambio, cuando la gerencia media deja de tener buena disposición (LECO=3 y LECO=4), el esfuerzo que insume el proyecto se incrementa.

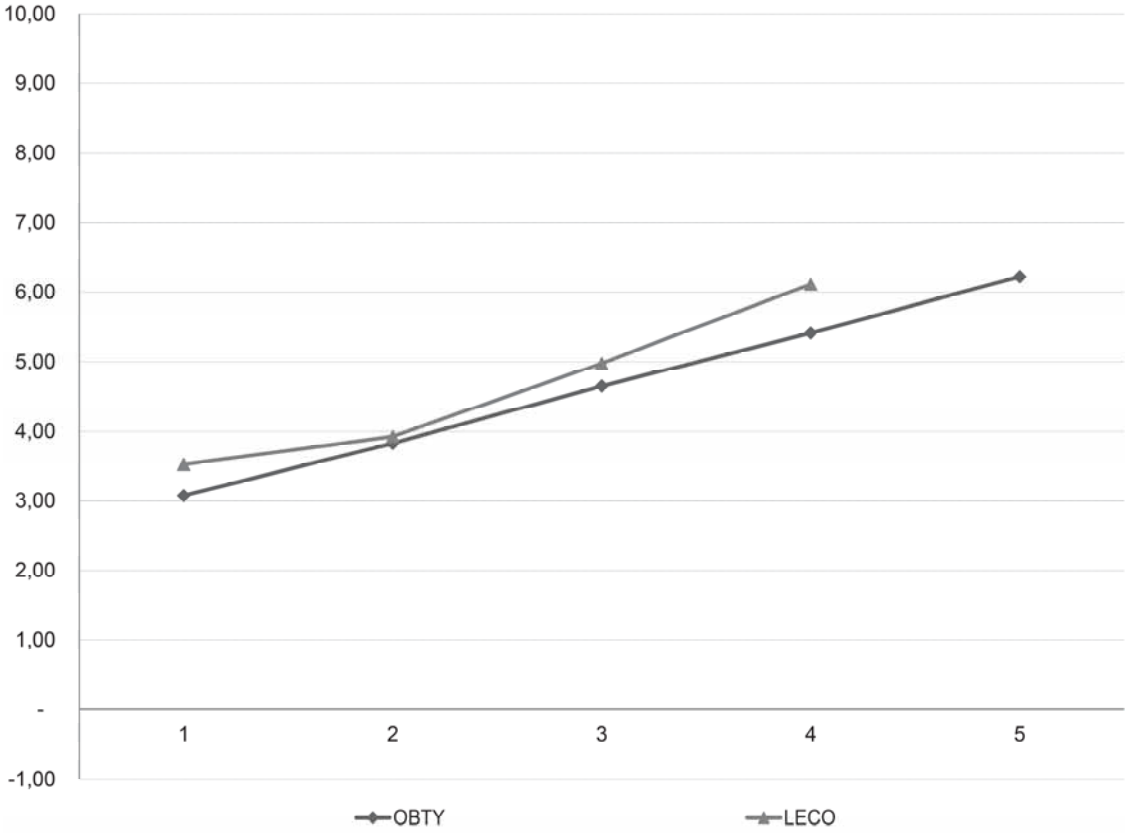


Fig. 4.16. Representación de la variación del esfuerzo estimado por la Fórmula Lineal según los factores de costo asociados al Proyecto.

- En el gráfico de la Figura 4.17 se indica la variación de los factores de costo relacionados a los *Datos Disponibles* donde se puede observar que:

  - Al aumentar el valor del factor de costo que trata sobre el nivel de conocimiento de los datos (KLDS), se produce un incremento del esfuerzo estimado; dado que este factor de costo aumenta a media que los repositorios de datos están menos documentados. Este crecimiento es más pronunciado en la medida en que es preciso consultar a expertos, al no existir ningún tipo de documentación disponible (KLDS=4).

    Asimismo, para este factor de costo se vuelve a notar el inconveniente identificado en la sección 4.3.3.2.1, en la cual el esfuerzo promedio de este factor de costo toma un valor negativo muy pequeño (-0,16 meses/hombre) para KLDS=1. A pesar de que este hecho constituye una inconsistencia, el mismo presenta algún sentido dado que al estar disponible la documentación de todos los repositorios y tablas, se facilita el trabajo de los ingenieros en el proyecto (sobre todo en la fase '*Comprensión de los Datos*' de la metodología CRISP-DM).

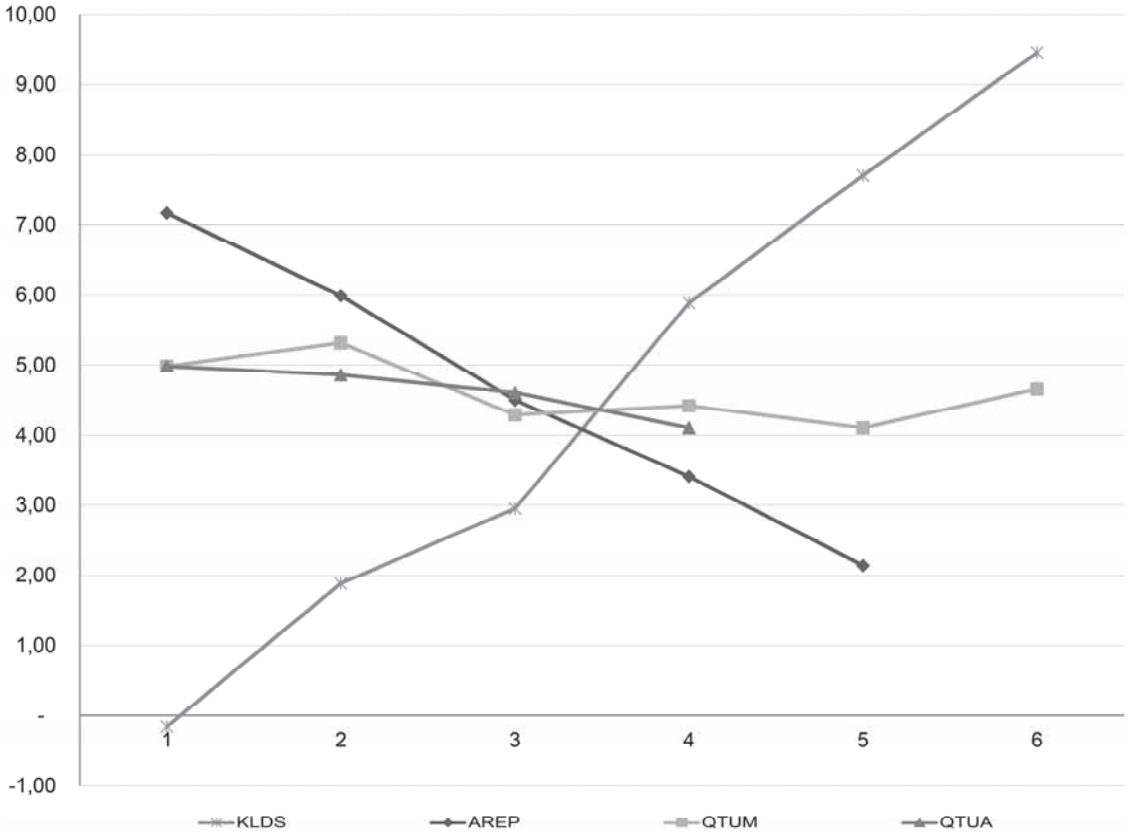


Fig. 4.17. Representación de la variación del esfuerzo estimado por la Fórmula Lineal según los factores de costo asociados a los Datos.

- En el caso de la cantidad y tipo de los repositorios de datos disponibles (AREP), se observa otra clase de inconsistencia. Aunque los valores estimados promedios son válidos (mayores a cero), el comportamiento de este factor presenta algunas diferencias con respecto a los anteriores, las cuales se indican a continuación.

  A medida que aumenta la cantidad de repositorios que se deben utilizar en el proyecto (crece el valor del factor de costo), el esfuerzo estimado disminuye. Por ejemplo, cuando se disponen de más de cinco repositorios (AREP=4 y AREP=5) el esfuerzo promedio es menor a 3,5 meses/hombre; mientras que cuando se dispone de un solo repositorio, se puede observar en la Figura 4.17 que el esfuerzo toma un valor mayor a 7 meses/hombre. Esto no es correcto, dado que a mayor cantidad de repositorios es normal que aumente el esfuerzo requerido (sobre todo en la fase '*Preparación de los Datos*').

- Para la cantidad de tuplas disponibles en la tabla principal (QTUM) el comportamiento de las estimaciones es muy variable por presentar crecimientos para los valores pares (QTUM=2, QTUM=4 y QTUM=6) y disminuciones para valores impares (QTUM=3 y QTUM=5). El mayor valor del esfuerzo promedio tiene lugar para QTUM=2 con 5,32

meses/hombre, siendo los siguientes máximos locales QTUM=1 (con 4,98 meses/hombre) y QTUM=6 (con 4,66 meses/hombre).

- El último factor de costo para este grupo es la cantidad de tuplas disponibles en las tablas auxiliares (QTUA), donde sucede algo similar al factor de costo *AREP*. En tal sentido, al aumentar la cantidad de tuplas en las tablas auxiliares, el esfuerzo disminuye. No obstante, se puede notar que la variación del esfuerzo promedio es mucho menor, habida cuenta que todos los valores se encuentran entre 4 y 5 meses/hombre.

- Finalmente, se presenta el gráfico que ilustra la variación de los siguientes factores de costo relacionados a los *Recursos Disponibles* (Figura 4.18), el cual se analiza a continuación:

  - Al observar el nivel de conocimiento y experiencia del equipo de trabajo (KEXT), se puede identificar una nueva inconsistencia: la misma consiste que al disminuir el nivel de conocimientos y experiencia del equipo de trabajo, el esfuerzo promedio estimado también disminuye en lugar de aumentar. Esto hecho no es correcto dado que es sabido que si se cuenta con un equipo de trabajo menos experimentado, es habitual que el esfuerzo requerido para realizar el proyecto sea mayor. En tal caso, el equipo va a necesitar mayor cantidad de tiempo para realizar las tareas que un equipo experimento podría realizar de manera más eficiente.

  - El factor que evalúa la funcionalidad de las herramientas disponibles (TOOL) presenta un crecimiento acentuado (de aproximadamente 40%), que es proporcional al aumento del valor del factor de costo. Esto es correcto dado que si se utilizan herramientas con menor cantidad de funciones para la preparación de datos, éstas deben ser realizadas por el equipo de trabajo en forma manual o desarrollando un software ad-hoc. Ambas circunstancias generan un aumento en el esfuerzo requerido para la realización del proyecto.

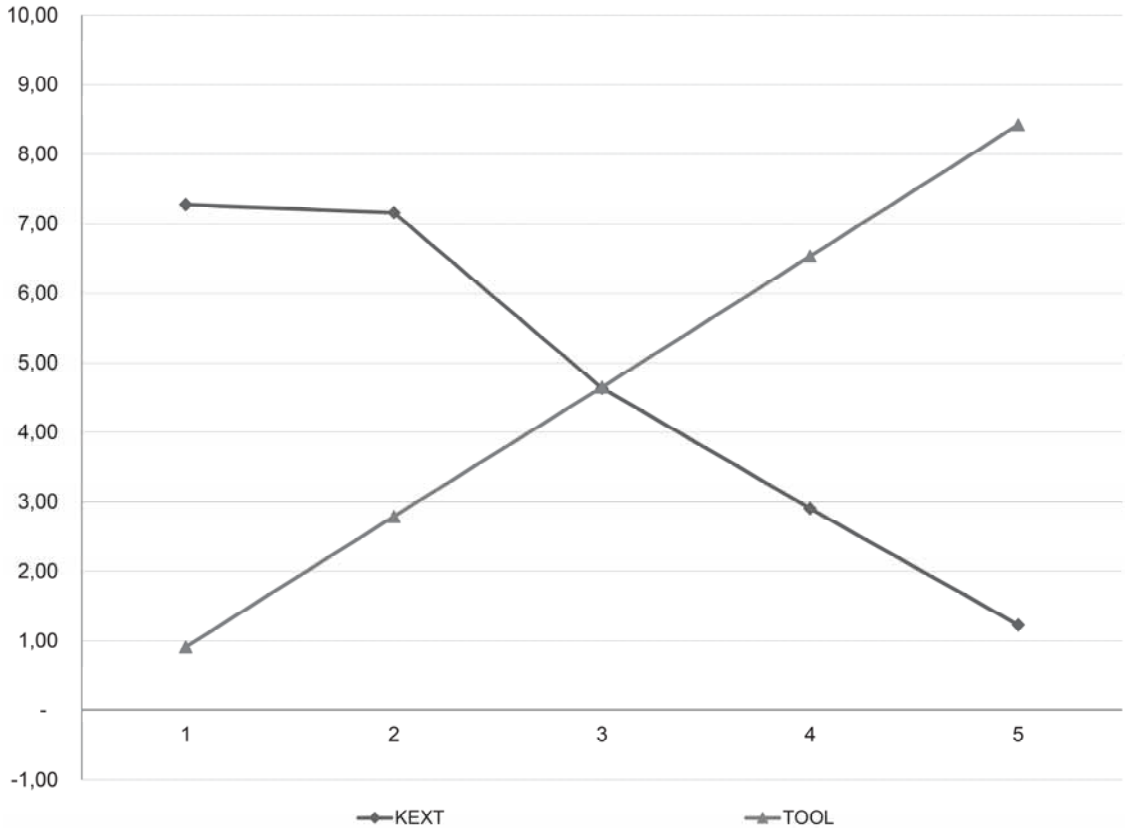


Fig. 4.18. Representación de la variación del esfuerzo estimado por la Fórmula Lineal según los factores de costo asociados a los Recursos.

4.3.3.2.3. Análisis Estadístico del Método Empírico para la Estimación de Esfuerzo

Como se ha mencionado anteriormente, debido a las inconsistencias detectadas en la fórmula lineal, el modelo de estimación propuesto incluye también un método empírico que es analizado a continuación. A tal efecto, se presentan gráficos similares a los utilizados en la sección anterior.

- Primero, en la Figura 4.19 se presenta el gráfico con la variación de los factores de costo relacionados al *Tipo del Proyecto* para el esfuerzo estimado por el método empírico donde se puede observar:

  - En el caso de tipo de objetivo de explotación de información (OBTY) se ve un crecimiento en la estimación de aproximadamente el 19% entre los dos primeros valores (entre OBTY=1 y OBTY=2), siendo luego el esfuerzo promedio constante. Esto significa que si el objetivo implica aplicar el proceso de descubrimiento de reglas de comportamiento el esfuerzo implicado menor será menor a los procesos de descubrimiento de grupos, ponderación de

atributos, descubrimiento de reglas de pertenencia a grupos y ponderación de reglas de comportamiento o de pertenencia a grupos. Aunque estos últimos procesos requieren un esfuerzo mayor al primero, el mismo es bastante similar entre sí.

- Para el grado de apoyo de los miembros de la organización (LECO) sucede algo similar: se nota un crecimiento significativo entre LECO=1 y LECO=2 (aproximadamente del 40%), y luego el esfuerzo se mantiene casi constante. Según este gráfico se puede indicar que mientras los directivos y el personal poseen buena disposición al proyecto se requiere menor esfuerzo, pero cuando el personal empieza a ser indiferente el esfuerzo requerido aumenta. Esto se debe a que, normalmente, es el personal operativo de la organización el que tiene los conocimientos necesarios sobre los datos a ser utilizados en el proyecto. En cambio, el no contar con el apoyo de la gerencia media, no genera una modificación significativa sobre el esfuerzo necesario.

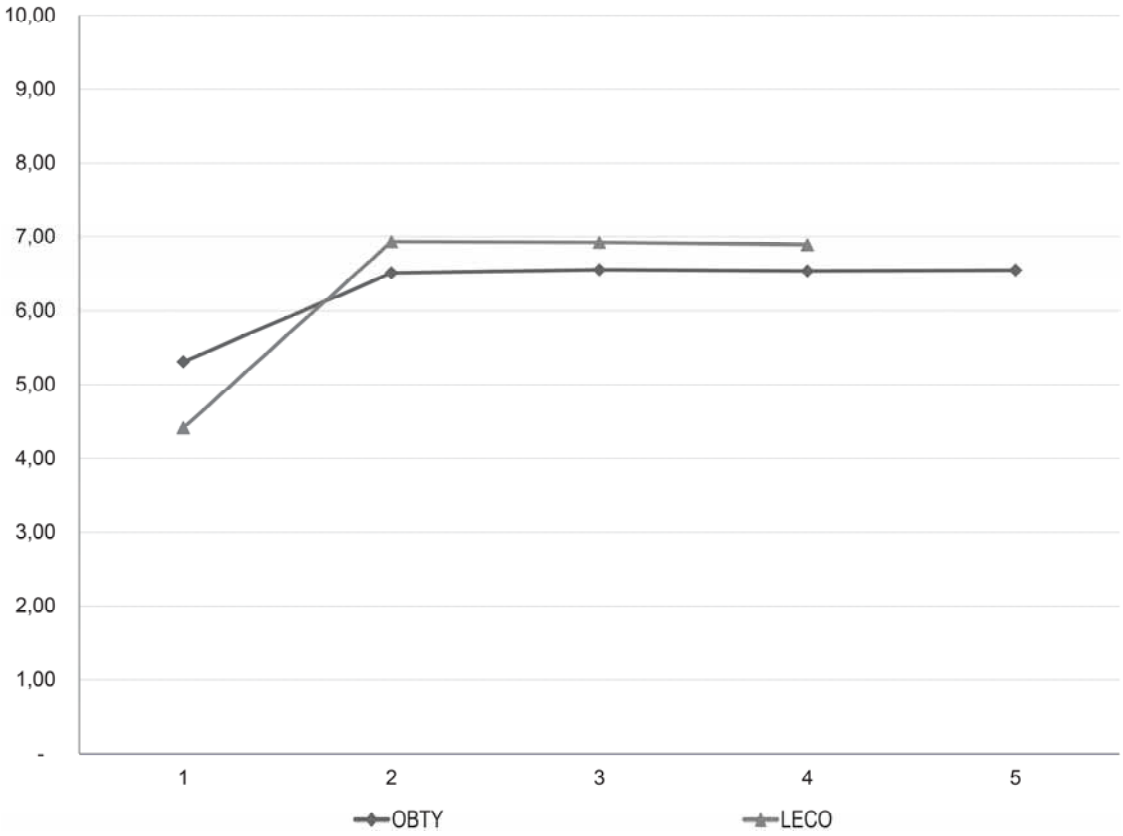


Fig. 4.19. Representación de la variación del esfuerzo estimado por el Método Empírico según los factores de costo asociados al Proyecto.

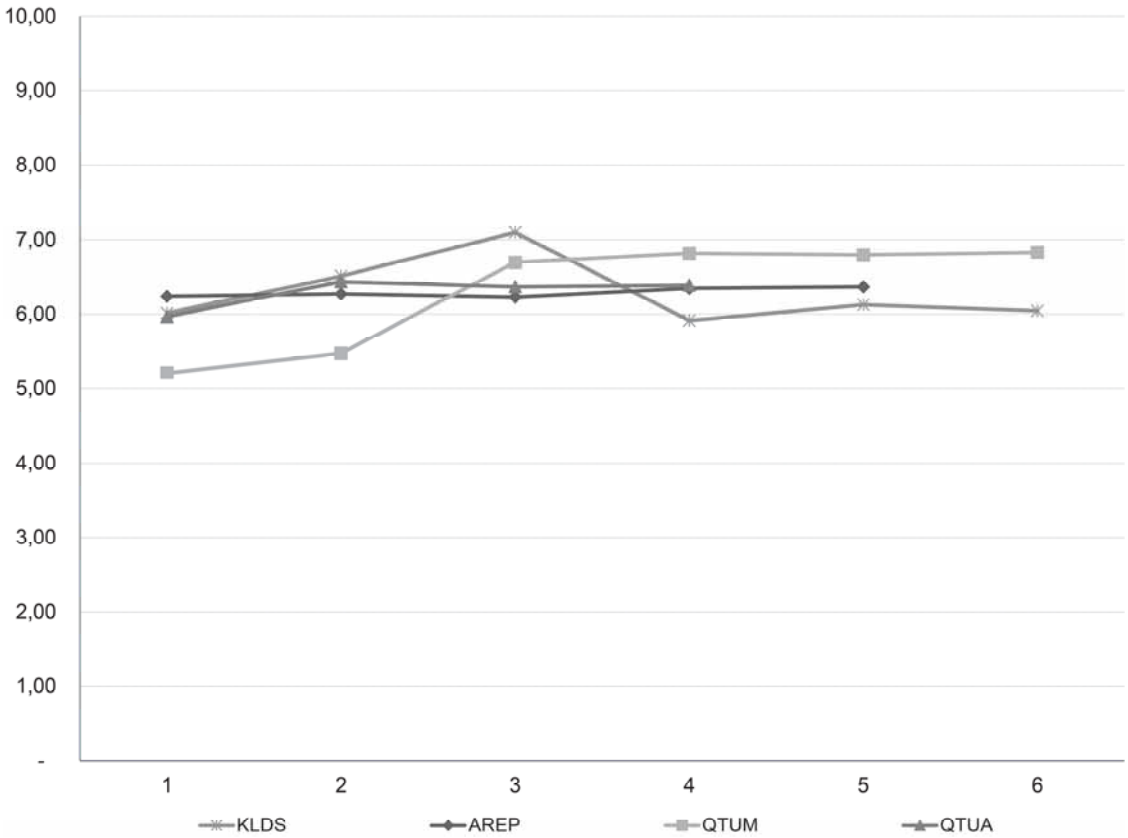


Fig. 4.20. Representación de la variación del esfuerzo estimado por el Método Empírico según los factores de costo asociados a los Datos.

- En el gráfico de la Figura 4.20 se indica la variación de los factores de costo relacionados a los *Datos Disponibles*:

  - En el caso de la cantidad y tipo de los repositorios de datos disponibles (AREP), el esfuerzo promedio no presenta modificaciones con respecto al aumento de la cantidad de repositorios. Por consiguiente, se puede decir que este factor de costo no tiene influencia significativa sobre el esfuerzo estimado.

  - Al observar la variación del conocimiento sobre los datos (KLDS), se puede notar que presenta un comportamiento bastante variable. El esfuerzo tiende a aumentar a medida que las fuentes de datos se encuentran menos documentadas (KLDS=1, KLDS=2 y KLDS=3), siendo el mayor esfuerzo detectado de 7,10 meses/hombre. En cambio, cuando las fuentes de datos no se encuentran documentadas y es necesario contactar a un experto (KLDS=4), el esfuerzo baja en casi un 19%. No obstante, si el experto no se encuentra disponible y no existe documentación (KLDS=5 y KLDS=6), el esfuerzo tiende nuevamente a crecer pero en menor medida (en el orden del 2%).

Este comportamiento se debe al hecho que recopilar la información necesaria para comprender los datos disponibles es más sencillo a través de una comunicación cara-a-cara con un experto que leyendo documentos por ser estos últimos más difíciles de comprender. Sin embargo, si no se puede contar con documentación ni un experto, el entendimiento de los datos debe llevarse a cabo a partir de los datos mismos usando técnicas y herramientas específicas.

- Para la cantidad de tuplas disponibles en la tabla principal (QTUM), el comportamiento de las estimaciones tiende a aumentar a medida que aumenta la cantidad de tuplas en la tabla principal. El mayor crecimiento (del 19%) se da cuando se superan los 1.000 tuplas (es decir entre QTUM=2 y QTUM=3) siendo luego la variación del esfuerzo bastante menor (en el orden del 1,5% aproximadamente)

- Por otra parte, para la cantidad de tuplas disponibles en las tablas auxiliares (QTUA), no existe gran variación del esfuerzo promedio por el aumento de las tuplas en dichas tablas. A pesar de que el esfuerzo tiende a crecer, la variación es muy pequeña (menos de medio mes/hombre) por lo que se puede decir que este factor de costo no parece afectar el esfuerzo estimado.

- Finalmente, en la Figura 4.21 se presenta el gráfico con la variación de los factores de costo relacionados a los *Recursos Disponibles* que es analizado a continuación:

  - Al observar el nivel de conocimiento y experiencia del equipo de trabajo (KEXT) se puede notar que, al disminuir el nivel de conocimientos y experiencia del equipo de trabajo, el esfuerzo promedio estimado sube. Como se ha mencionado anteriormente, si se cuenta con un equipo de trabajo menos experimentado, normalmente el esfuerzo requerido para llevar a cabo el proyecto va a ser mayor. Esto se debe a que normalmente un equipo con menos experiencia va a necesitar más tiempo para realizar las mismas tareas que uno experimentado. Pero este crecimiento no es lineal. Para casos donde se busca obtener los mismos objetivos con datos o tipos de organización diferentes (KEXT=2 y KEXT=3), se requiere el mismo esfuerzo promedio. En cambio, si los objetivos son similares pero tanto los datos como el tipo de organización son diferentes (KEXT=4), el esfuerzo aumenta alrededor de un 12%. Por último, si no se cuenta con ninguna de experiencia (objetivos, tipo

de organización y datos diferentes en KEXT=5), el esfuerzo es mucho mayor, creciendo en el orden del 23%

- Con el factor asociado a la funcionalidad de las herramientas disponibles (TOOL) se observa un crecimiento acentuado (de aproximadamente 48%), sólo cuando la herramienta no incluye funciones para realizar el formateo de los datos (es decir entre TOOL=3 y TOOL=4). En los otros casos, entre TOOL=1, TOOL=2 y TOOL=3 por una parte, y entre TOOL=4 y TOOL=5 por la otra, no se presenta crecimiento siendo el esfuerzo promedio similar. Esto significa que es más importante contar con una herramienta que asista al ingeniero en la preparación de los datos debido a que permite reducir significativamente el esfuerzo requerido en el proyecto.

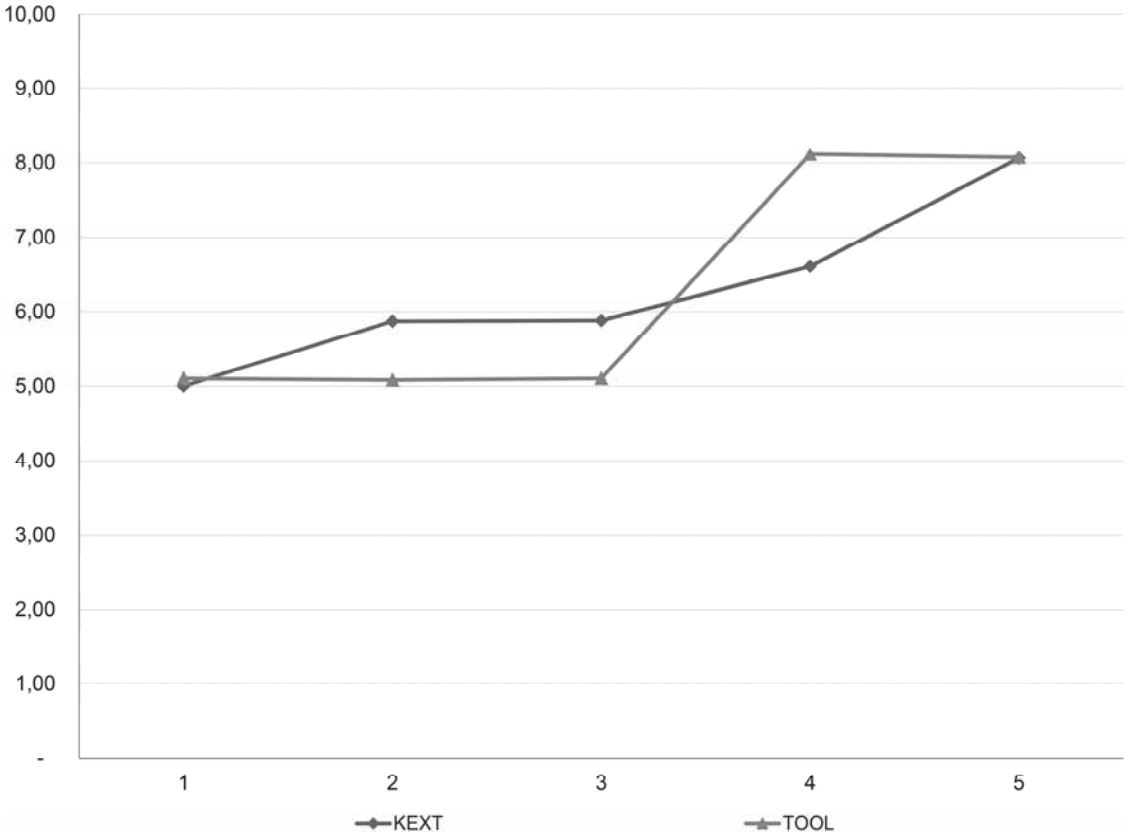


Fig. 4.21. Representación de la variación del esfuerzo estimado por el Método Empírico según los factores de costo asociados a los Recursos.

4.3.3.2.4. Conclusiones del Análisis Estadístico para el Modelo de Estimación de Esfuerzo

De los estudios estadísticos llevados a cabo sobre el modelo de estimación de esfuerzo se puede concluir que la fórmula de estimación lineal tiende a generar un comportamiento inconsistente para

ciertas combinaciones de los valores de los factores de costo. Como se ha indicado en la sección 4.3.3.2.1, para ciertos tipos de proyectos se han obtenido estimaciones inválidas (con valores negativos). A partir de la consulta a expertos en el campo de la Explotación de Información, se ha determinado que estas combinaciones de valores no suelen generarse en la realidad para proyectos llevados a cabo en PyMEs. De hecho, de todos los proyectos generados en forma pseudo-aleatoria, este problema aplica a menos del 20%. En consecuencia, se decide utilizar la fórmula para ser validada en el siguiente capítulo de esta tesis.

De manera alternativa, el comportamiento del método empírico de estimación no presenta dicho problema. La única crítica que se le podría hacer es que método es más conservador por mantener las estimaciones entre medio año/hombre y un año/hombre, mientras que la fórmula lineal genera estimaciones cercanas a los 2 años/hombre.

Para ambos métodos es posible destacar las siguientes factores de costo como determinantes para calcular el esfuerzo de un proyecto: el grado de apoyo de los miembros de la organización, el nivel de conocimiento sobre los datos, la cantidad de tuplas disponibles en la tabla principal, la funcionalidad de las herramientas disponibles y nivel de conocimiento y, por supuesto, el nivel de experiencia del equipo de trabajo. Por consiguiente, estas características deben ser cuidadosamente evaluadas para obtener una estimación confiable; y, de esta manera, el Ingeniero de Explotación de Información contar con información precisa al momento de planificar actividades, personas y recursos necesarios para llevar a cabo del proyecto.

5. VALIDACIÓN

En este capítulo se presentan los resultados obtenidos de la validación de los dos modelos propuestos en el capítulo anterior. A tal efecto, en primer término se realiza una introducción sobre los aspectos generales de la validación efectuada (sección 5.1). En segundo término, se lleva a cabo la validación del modelo para la Evaluación de la Viabilidad de Proyectos pequeños y medianos de Explotación de Información (sección 5.2), donde se indican los datos utilizados (sección 5.2.1), se presentan los resultados obtenidos por un estudio estadístico (sección 5.2.2) y por la prueba de Wilcoxon (sección 5.2.3); para de esta manera, llegar a establecer las conclusiones preliminares acerca de este modelo (sección 5.2.4). En tercer término, se hace uso de una estructura similar en la validación del modelo para la Estimación de Esfuerzo de Proyectos pequeños y medianos de Explotación de Información (sección 5.3); se indican los datos utilizados (sección 5.3.1), se presentan los resultados del análisis estadístico (sección 5.3.2) y de la prueba de Wilcoxon (sección 5.3.3); y de esta forma establecer las conclusiones preliminares correspondientes a este modelo (sección 5.3.4).

5.1. INTRODUCCIÓN

La validación de los modelos propuestos en el capítulo anterior se lleva a cabo mediante la comparación de la información recolectada para treinta y siete proyectos de Explotación de Información reales, con los resultados de la aplicación de cada uno de los modelos propuestos. A partir del análisis de esta comparación, se determina si los modelos propuestos son confiables en el ámbito de las Pequeñas y Medianas Empresas (PyMEs).

Los proyectos utilizados en esta validación fueron suministrados por investigadores del Grupo de Investigación en Sistemas de Información del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús (GISI-DDPyT-UNLa), investigadores del Grupo de Estudio en Metodologías de Ingeniería de Software de la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional (GEMIS-FRBA-UTN), e investigadores del Grupo de Investigación en Explotación de Información en el Laboratorio de Informática Aplicada de la Universidad Nacional de Río Negro (GIEdI-UNRN). Tal como se mencionó en la sección 4.1, todos estos proyectos fueron desarrollados mediante la aplicación de la metodología CRISP-DM [Chapman *et al.*, 2000]. Por consiguiente, las conclusiones de esta validación se pueden considerar confiables sólo para proyectos de Explotación de Información que se desarrollan bajo esta metodología así como a partir del Modelo de Proceso para Proyectos de Explotación de Información [Vanrell *et al.*, 2010; 2012]; por ser este último una adaptación de CRISP-DM.

5.2. VALIDACIÓN DEL MODELO PARA EVALUACIÓN DE LA VIABILIDAD DE PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

En esta sección se presenta la validación del modelo propuesto para evaluar la viabilidad de un proyecto de Explotación de Información dentro del ámbito de una PyME. La sección se encuentra estructurada de la siguiente forma: los datos utilizados para llevar a cabo esta validación (sección 5.2.1), la realización de un análisis estadístico sobre dichos datos (sección 5.2.2), la aplicación de la prueba de Wilcoxon (sección 5.2.3); y la presentación de las conclusiones preliminares acerca de este modelo (5.2.4).

5.2.1. Datos utilizados en la Validación del Modelo para la Evaluación de la Viabilidad

En la validación del modelo para la Evaluación de la Viabilidad se utilizan todos los proyectos recolectados. De los treinta y siete proyectos disponibles, los primero treinta y dos (es decir, P1 a P32 inclusive) han sido desarrollado en forma completa y sus resultados pudieron ser utilizados por la organización; por tal razón, se considera estos proyectos como finalizados satisfactoriamente. Por otra parte, los cinco proyectos restantes (P33 a P37), fueron cancelados antes de su finalización debido a problemas identificados durante su proceso de desarrollo.

Tal como se ha mencionado en la sección anterior, estos proyectos son utilizados para llevar a cabo un estudio comparativo entre los valores de las dimensiones obtenidos por el modelo (plausibilidad, adecuación y éxito) y una valoración de carácter subjetiva de dichas dimensiones (en una escala de 1 a 10) proporcionada por investigadores considerados como expertos en proyectos de Explotación de Información.

Con los tres valores suministrados por los expertos, se calcula su promedio a los efectos de obtener el valor de la ‘Valoración de la Viabilidad del proyecto’; el cual se correspondería con la ‘Viabilidad Global’ (EV) que se calcula por medio del modelo propuesto. Los resultados de este proceso de valoración se encuentran disponibles también en el Anexo B de este trabajo.

Por otra parte, con independencia de la valoración realizada por los investigadores, a cada proyecto se le aplica los cinco pasos del modelo de Viabilidad propuesto. Estos resultados se encuentran disponibles en el anexo C de este trabajo.

A modo de resumen, en la Tabla 5.1 se presentan para cada proyecto, los valores asignados por los investigadores para cada dimensión (con su promedio correspondiente) junto con los resultados de la aplicación del modelo propuesto.

#	Estado Final del Proyecto	Valoración indicada por los Investigadores				Valores calculados por el Modelo de Viabilidad			
		Plausibilidad	Adecuación	Éxito	Viabilidad Global	Plausibilidad	Adecuación	Éxito	Viabilidad Global
P1	*Finalizado Satisfactoriamente*	*8*	*7*	*4*	*6,33*	*7,20*	*6,11*	*5,25*	*6,27*
P2	*Finalizado Satisfactoriamente*	*7*	*6*	*5*	*6,00*	*6,87*	*5,07*	*5,25*	*5,77*
P3	*Finalizado Satisfactoriamente*	*8*	*5*	*6*	*6,33*	*5,90*	*5,67*	*5,31*	*5,65*
P4	*Finalizado Satisfactoriamente*	*6*	*6*	*4*	*5,33*	*5,12*	*6,95*	*4,12*	*5,51*
P5	*Finalizado Satisfactoriamente*	*6*	*8*	*7*	*7,00*	*5,12*	*7,82*	*6,81*	*6,56*
P6	*Finalizado Satisfactoriamente*	*6*	*5*	*5*	*5,33*	*5,45*	*5,61*	*5,25*	*5,45*
P7	*Finalizado Satisfactoriamente*	*5*	*5*	*5*	*5,00*	*5,45*	*5,56*	*5,42*	*5,48*
P8	*Finalizado Satisfactoriamente*	*6*	*5*	*6*	*5,67*	*6,45*	*5,80*	*5,18*	*5,87*
P9	*Finalizado Satisfactoriamente*	*7*	*6*	*6*	*6,33*	*7,20*	*5,61*	*5,57*	*6,18*
P10	*Finalizado Satisfactoriamente*	*6*	*5*	*6*	*5,67*	*5,85*	*5,34*	*5,57*	*5,59*
P11	*Finalizado Satisfactoriamente*	*8*	*5*	*6*	*6,33*	*6,22*	*6,56*	*5,42*	*6,14*
P12	*Finalizado Satisfactoriamente*	*7*	*8*	*7*	*7,33*	*7,67*	*7,35*	*6,45*	*7,22*
P13	*Finalizado Satisfactoriamente*	*7*	*5*	*6*	*6,00*	*5,93*	*5,09*	*7,05*	*5,93*
P14	*Finalizado Satisfactoriamente*	*7*	*7*	*6*	*6,67*	*6,20*	*6,59*	*5,69*	*6,20*
P15	*Finalizado Satisfactoriamente*	*9*	*7*	*8*	*8,00*	*8,72*	*6,89*	*7,66*	*7,77*
P16	*Finalizado Satisfactoriamente*	*7*	*6*	*5*	*6,00*	*6,45*	*6,43*	*5,64*	*6,22*
P17	*Finalizado Satisfactoriamente*	*6*	*5*	*5*	*5,33*	*6,14*	*5,83*	*5,42*	*5,83*
P18	*Finalizado Satisfactoriamente*	*5*	*5*	*6*	*5,33*	*6,00*	*5,31*	*5,42*	*5,59*
P19	*Finalizado Satisfactoriamente*	*8*	*7*	*7*	*7,33*	*7,01*	*6,89*	*5,58*	*6,58*
P20	*Finalizado Satisfactoriamente*	*9*	*7*	*5*	*7,00*	*8,24*	*6,75*	*5,52*	*6,96*
P21	*Finalizado Satisfactoriamente*	*8*	*6*	*5*	*6,33*	*8,05*	*6,45*	*5,25*	*6,70*
P22	*Finalizado Satisfactoriamente*	*7*	*6*	*6*	*6,33*	*6,45*	*5,81*	*6,54*	*6,24*
P23	*Finalizado Satisfactoriamente*	*7*	*7*	*8*	*7,33*	*6,87*	*5,20*	*5,96*	*6,01*
P24	*Finalizado Satisfactoriamente*	*7*	*8*	*5*	*6,67*	*8,05*	*6,76*	*5,81*	*6,97*
P25	*Finalizado Satisfactoriamente*	*5*	*7*	*5*	*5,67*	*6,00*	*6,76*	*5,00*	*6,00*
P26	*Finalizado Satisfactoriamente*	*8*	*8*	*8*	*8,00*	*6,55*	*7,00*	*5,01*	*6,29*
P27	*Finalizado Satisfactoriamente*	*8*	*6*	*7*	*7,00*	*6,00*	*6,70*	*6,54*	*6,40*
P28	*Finalizado Satisfactoriamente*	*8*	*6*	*7*	*7,00*	*6,39*	*5,58*	*5,47*	*5,85*
P29	*Finalizado Satisfactoriamente*	*7*	*5*	*7*	*6,33*	*7,64*	*6,27*	*6,45*	*6,82*
P30	*Finalizado Satisfactoriamente*	*8*	*8*	*6*	*7,33*	*6,87*	*5,90*	*4,97*	*6,00*
P31	*Finalizado Satisfactoriamente*	*7*	*6*	*8*	*7,00*	*6,52*	*6,39*	*6,54*	*6,48*
P32	*Finalizado Satisfactoriamente*	*7*	*7*	*8*	*7,33*	*6,60*	*6,39*	*6,20*	*6,42*
P33	*Cancelado*	*3*	*4*	*3*	*3,33*	*4,49*	*4,77*	*4,99*	*4,73*
P34	*Cancelado*	*4*	*5*	*2*	*3,67*	*4,36*	*4,62*	*2,64*	*3,99*
P35	*Cancelado*	*3*	*4*	*3*	*3,33*	*4,66*	*5,34*	*3,25*	*4,52*
P36	*Cancelado*	*5*	*3*	*2*	*3,33*	*4,66*	*3,46*	*4,21*	*4,10*
P37	*Cancelado*	*4*	*2*	*1*	*2,33*	*4,63*	*2,81*	*3,01*	*3,52*

Tabla 5.1. Datos de los proyectos usados en la validación del modelo de viabilidad.

A partir de la observación de la Tabla 5.1, es posible determinar que, en líneas generales, el modelo logra estimar correctamente si el proyecto ha finalizado satisfactoriamente o si ha sido cancelado en casi todos los casos. Tal como se definió en la sección 4.2.2.2, el criterio del modelo que permite establecer si un proyecto es viable, consiste en lo siguiente: el resultado de todas las dimensiones debe ser superior al intervalo correspondiente al valor lingüístico '*regular*' (lo cual es equivalente al valor numérico 5), y la valoración global de la viabilidad proyecto (EV) debe ser mayor a 5. Asimismo, cabe recordar que un mayor valor de cada una de estas dimensiones se corresponde con una dimensión mejor valorada.

En función de lo expresado anteriormente, los proyectos que no se consideran viables son: P4, P30, P33, P34, P35, P36 y P37. Los dos primeros no son viables, dado que el valor obtenido para la dimensión éxito es menor que 5 (aunque con un valor muy cercano), mientras que los restantes proyectos no superan el valor mínimo en varias de sus dimensiones. En tal sentido, en los últimos cinco proyectos (P33, P34, P35, P36 y P37) la evaluación es correcta por tratarse de proyectos que fueron cancelados antes de su terminación; mientras que para los proyectos P4 y P30, la evaluación realizada por el modelo es errónea, dado que estos proyectos finalizaron de manera satisfactoria. En función de lo comunicado por sus líderes de proyecto, estos dos proyectos mencionados han presentado muchos problemas durante el desarrollo, y han finalizado de forma satisfactoria sólo en virtud de un esfuerzo adicional realizado por los participantes del equipo de trabajo. Este hecho permite indicar que el modelo ha sido capaz de predecir correctamente las escasas posibilidades de finalizar con éxito que tenían dichos proyectos.

5.2.2. Análisis Estadístico del Modelo para la Evaluación de la Viabilidad

A partir de los datos indicados en la Tabla 5.1, en esta sección se lleva a cabo un análisis estadístico del modelo de viabilidad. Dicho análisis consiste en la comparación de la valoración indicada por los investigadores con el valor calculado por el modelo para cada dimensión. En tal sentido, el análisis se realiza tomando cada dimensión (plausibilidad, adecuación, éxito y viabilidad global) de forma independiente.

- Análisis de la *Plausibilidad:*

  La primera dimensión analizada es la plausibilidad cuyos resultados estadísticos se presentan en la Tabla 5.2. En dicha tabla se indican el valor mínimo, valor máximo, valor promedio y valor de varianza tanto de los valores suministrados por los investigadores expertos así como los generados por el modelo. Asimismo, para facilitar la interpretación, estos valores se reflejan en

el gráfico Boxplot [Turkey, 1977] de la Figura 5.1. Se recuerda que este tipo de gráfico permite representar en una única figura los datos correspondientes a los límites superior e inferior (valores máximo y mínimo) mediante una línea más fina, su región crítica que está comprendida entre el desvío máximo (media más la desviación estándar) y mínimo (media menos la desviación estándar) mediante una línea más gruesa, y el valor medio (o promedio) de las estimaciones obtenidas mediante una línea horizontal.

Parámetro Estadístico	**Plausibilidad**	
	Valoración indicada por los Investigadores	**Valor calculado por el Modelo de Viabilidad**
Valor Mínimo	*3,00*	*4,36*
Valor Máximo	*9,00*	*8,72*
Valor Promedio	*6,59*	*6,32*
Valor de Varianza	*2,30*	*1,20*

Tabla 5.2. Resultados estadísticos para la dimensión Plausibilidad.

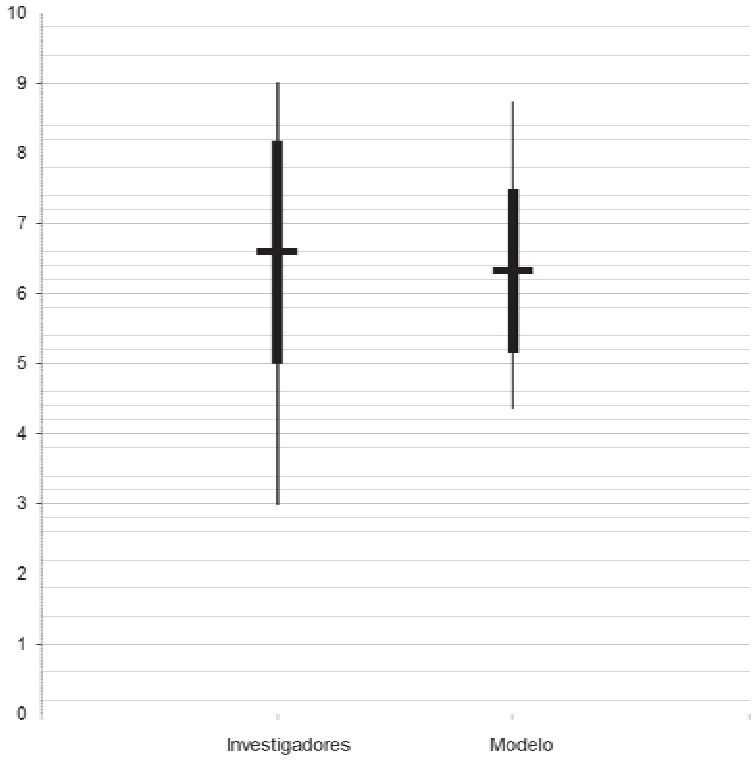


Fig. 5.1. Gráfico boxplot de la dimensión Plausibilidad.

Como se puede observar en la Figura 5.1, los valores calculados por el modelo para la plausibilidad son similares a los indicados por los investigadores. A pesar de que la varianza de los valores asignados por los investigadores es un poco mayor, el modelo genera valores muy similares dado que su región crítica (comprendida entre el valor de desvío mínimo y máximo) se

encuentra contenida dentro de la otra. En virtud de lo expresado en la Tabla 5.2, se observa que la mayor diferencia se genera entre el valor mínimo indicado por los investigadores y el valor mínimo calculado por el modelo; en tal sentido, existe una diferencia de 1,36 unidades. En cambio, para el promedio y los valores máximos, la diferencia es aproximadamente menor a 0,30 unidades. Por consiguiente, en este caso se puede decir que el modelo es más conservador con respecto a lo que expresan los investigadores, debido a que los valores calculados por el modelo son más acotados y se encuentran contenidos dentro de los valores asignados por los investigadores.

Con el objetivo de llevar a cabo un análisis más detallado, se procede a comparar los valores indicados por los investigadores con los calculados por el modelo para cada proyecto en el gráfico que se muestra en la Figura 5.2. De esta manera, los valores suministrados por los investigadores se representan en forma de línea y los generados por el modelo en forma de barras.

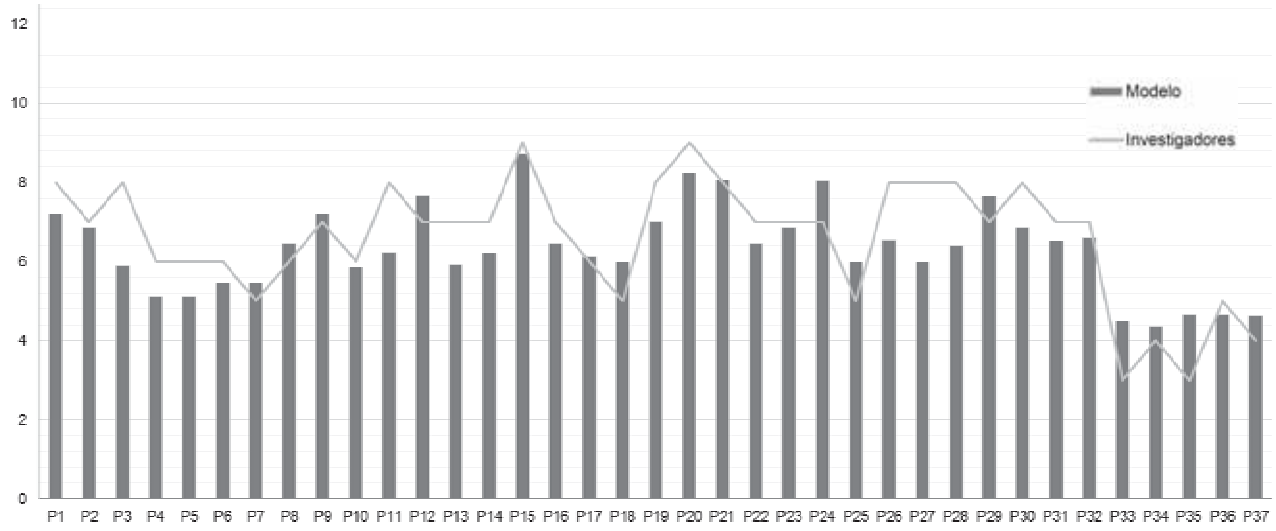


Fig. 5.2. Gráfico de comparación de los valores de la Plausibilidad.

En dicho gráfico se puede notar que en general no hay diferencias significativas entre los dos tipos de valores. El mayor error detectado corresponde al proyecto P3 (con una subestimación generada por el modelo de 2,10 unidades), mientras que treinta y tres de los proyectos (es decir, 89% de los proyectos) tienen un error relativo menor al 25%. Además de P3, los otros tres proyectos que sobrepasan ese porcentaje de error son P27 (con una subestimación de exactamente 2 unidades), P33 (con una sobreestimación generada por el modelo de aproximadamente 1,50 unidades) y P35 (con una sobreestimación de aproximadamente 1,7 unidades). De manera alternativa, el 51% de los proyectos tienen un error relativo menor al

10%, cumpliéndose esto para diecinueve de los proyectos analizados (es decir para P2, P6, P7, P8, P9, P10, P12, P15, P16, P17, P20, P21, P22, P23, P29, P31, P32, P34 y P36).

- Análisis de la *Adecuación*:

  En segundo término, se analiza de forma similar la dimensión adecuación a partir de los valores que se indican en la Tabla 5.3 y el gráfico Boxplot de la Figura 5.3.

Parámetro Estadístico	**Adecuación**	
	Valoración indicada por los Investigadores	**Valor calculado por el Modelo de Viabilidad**
Valor Mínimo	*2,00*	*2,81*
Valor Máximo	*8,00*	*7,82*
Valor Promedio	*5,89*	*5,93*
Valor de Varianza	*1,99*	*1,02*

Tabla 5.3. Resultados estadísticos para la dimensión Adecuación.

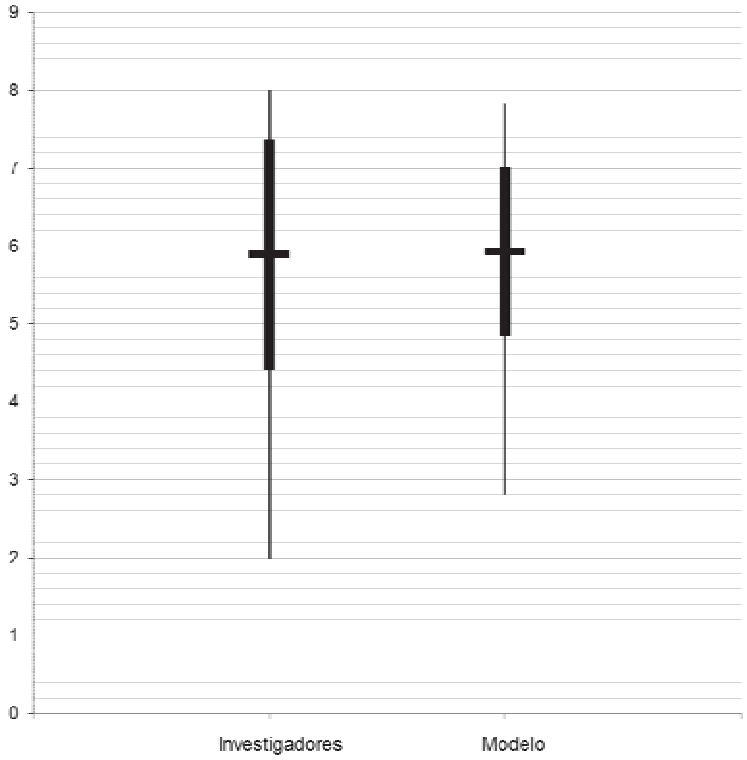


Fig. 5.3. Gráfico boxplot de la dimensión Adecuación.

Para esta dimensión se puede observar nuevamente que el modelo genera valores muy similares a los indicados por los investigadores. En este caso, la varianza es muy cercana debido a que el modelo genera valores muy similares: mientras que los valores mínimos poseen una diferencia de 0,81 unidades, para los máximos es menor a 0,20 y el promedio

menor a 0,05. Tal como sucede con la dimensión de plausibilidad, para la adecuación, el modelo también tiende a ser algo conservador; aunque en menor medida.

Asimismo, como parte de un análisis más detallado, se comparan los valores para cada proyecto tal como se puede ver en la Figura 5.4. En este caso tampoco se pueden notar diferencias significativas entre las dos clases de valores. Salvo cuatro proyectos (P23, P30, P33 y P35), el resto tienen un error relativo menor al 25%, siendo el mayor error de 2,10 unidades. Por otra parte, el 57% de los proyectos tienen un error relativo menor al 10%, cumpliéndose esto para veintiuno de los mismos (es decir para P3, P5, P9, P10, P12, P13, P14, P15, P16, P18, P19, P20, P21, P22, P25, P27, P28, P31, P32, P34 y P36).

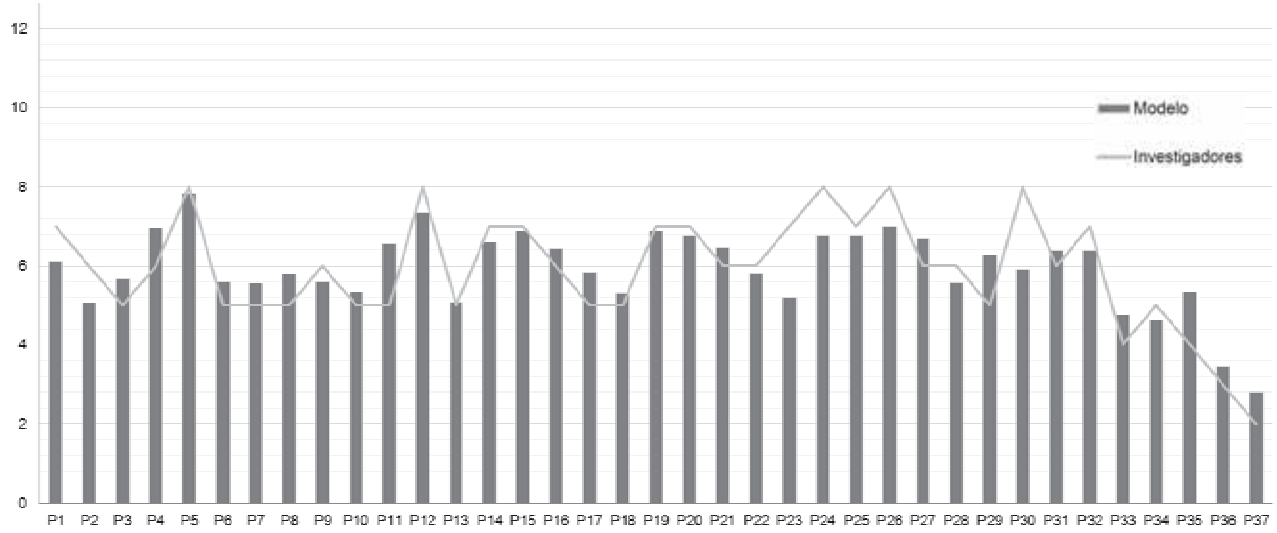


Fig. 5.4. Gráfico de comparación de los valores de la Adecuación.

- Análisis del *Éxito*:

  En este caso también se indican los resultados del análisis en la Tabla 5.4 y la Figura 5.5.

Parámetro Estadístico	**Éxito**	
	Valoración indicada por los Investigadores	**Valor calculado por el Modelo de Viabilidad**
Valor Mínimo	*1,00*	*2,64*
Valor Máximo	*8,00*	*7,66*
Valor Promedio	*5,57*	*5,44*
Valor de Varianza	*3,09*	*1,08*

Tabla 5.4. Resultados estadísticos para la dimensión Éxito.

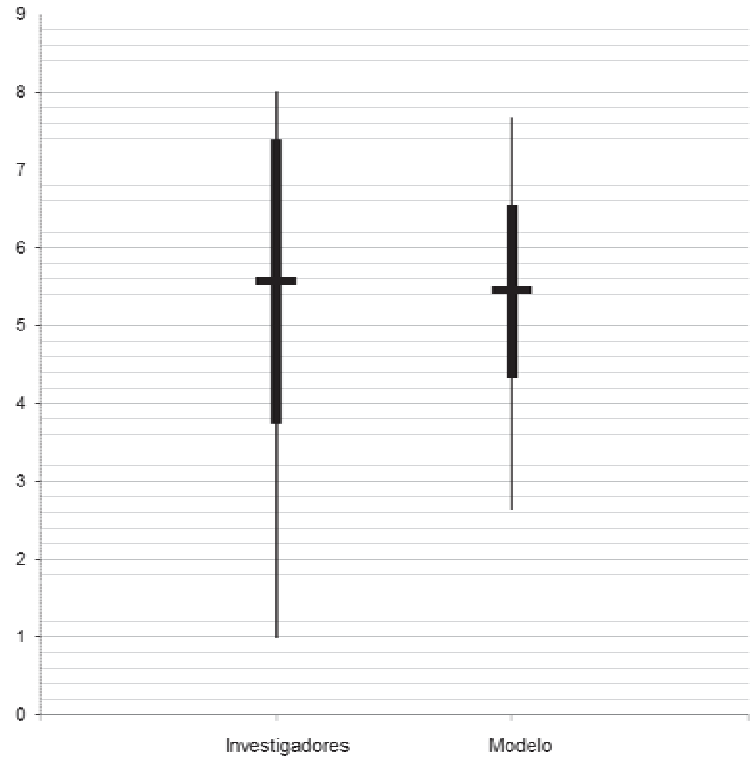


Fig. 5.5. Gráfico boxplot de la dimensión Éxito.

Para esta dimensión se observa un comportamiento similar al caso de plausibilidad debido a que la mayor diferencia se encuentra entre los valores mínimos. No obstante, cabe señalar que el error es un poco superior entre la valoración indicada por los investigadores y los valores calculados según el modelo en los extremos (valores mínimo y máximo); mientras que para la media, el error es inferior. Por otra parte, en esta dimensión se puede apreciar que el modelo es muy conservador con respecto a lo definido por los investigadores, por ser su varianza mucho más pequeña.

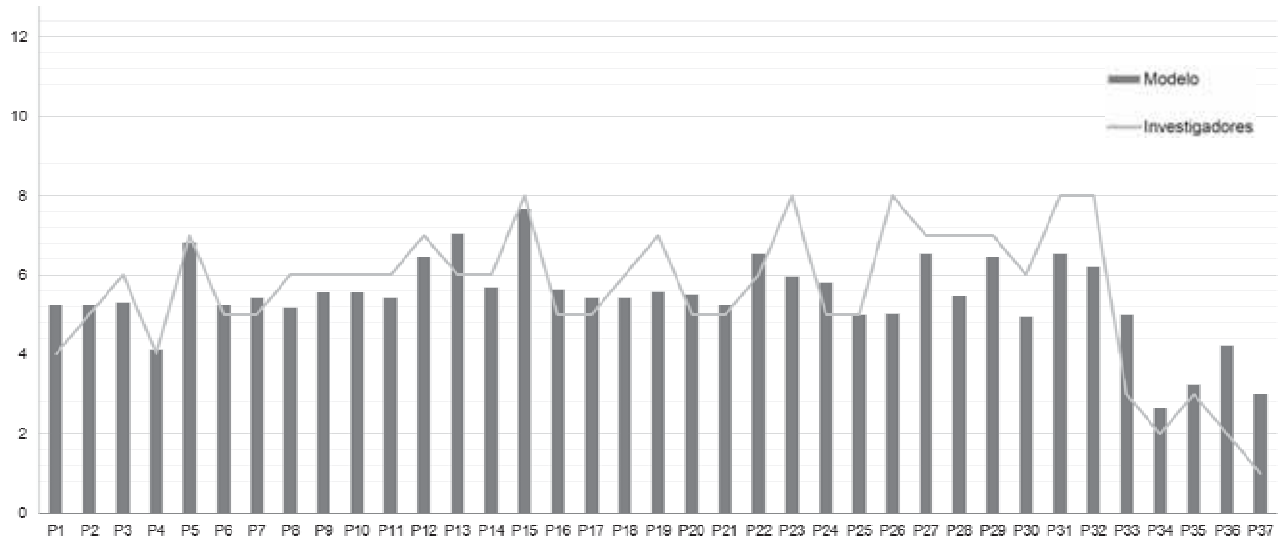


Fig. 5.6. Gráfico de comparación de los valores de la dimensión Éxito.

Sin embargo, al comparar los resultados por proyecto (Figura 5.6), se puede observar que sólo en los proyectos P23, P26, P32, P33, P36 y P37 el error relativo es mayor al 25% (con un máximo de casi 3 unidades). En cambio, para veintiuno de los proyectos un error menor al 10% (o sea para P2, P3, P4, P5, P6, P7, P9, P10, P11, P12, P14, P15, P16, P17, P20, P21, P22, P25, P27, P29 y P35) por lo que la efectividad es similar que la obtenida en las dos dimensiones anteriores (aproximadamente del 57%).

- Análisis de la *Viabilidad Global del Proyecto*:

  A pesar de ser la última dimensión analizada, ésta es la más significativa por ser calculado en función de las otras tres dimensiones. La Viabilidad Global sintetiza los resultados parciales obtenidos para las otras dimensiones; permitiendo, de esta manera, determinar si el proyecto es viable o no. Cabe recordar que la Viabilidad Global se calcula utilizando una media aritmética ponderada de las dimensiones de plausibilidad, adecuación y éxito, tal como se define en el proceso de la sección 4.2.2.2 del capítulo anterior. Los resultados estadísticos generales de esta última dimensión se expresan en la Tabla 5.5 y la Figura 5.7.

Parámetro Estadístico	**Viabilidad Global**	
	Valoración indicada por los Investigadores	**Valor calculado por el Modelo de Viabilidad**
Valor Mínimo	*2,33*	*3,52*
Valor Máximo	*8,00*	*7,77*
Valor Promedio	*6,02*	*5,94*
Valor de Varianza	*1,86*	*0,78*

Tabla 5.5. Resultados estadísticos para la dimensión Viabilidad Global.

En función a los valores expresados en la Tabla 5.5, se infiere que la diferencia entre los valores asignados por los investigadores y los calculados por el modelo es muy pequeña. En este sentido, se puede observar que la mayor diferencia (de aproximadamente 1,20 unidades) tiene lugar para el parámetro estadístico correspondiente a los valores mínimos; mientras que para los valores máximos y el promedio la diferencia es menor a 0,25 unidades. Esto hecho es muy significativo, dado que es consistente con la naturaleza conservadora del modelo; en virtud de que el mismo proporciona valores muy bajos, lo cual se detecta en los análisis realizados para las otras tres dimensiones.

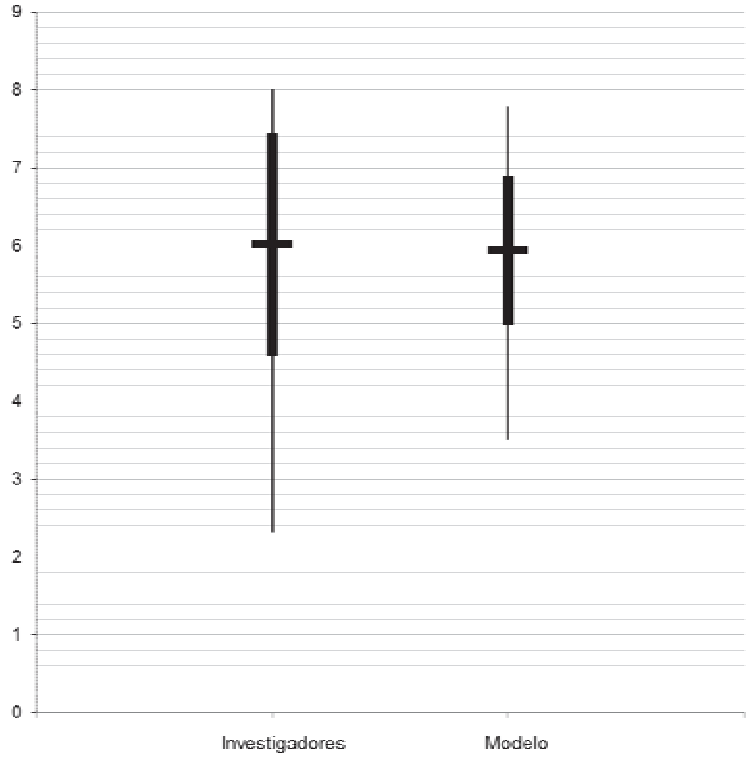


Fig. 5.7. Gráfico boxplot de la Viabilidad Global.

Al comparar los resultados proyecto a proyecto (Figura 5.8), se observa que en aquellos proyectos considerados como viables (P1 a P32), los valores calculados del modelo son similares a los indicados por los investigadores; habida cuenta de que se tiene un error relativo menor al 25%. Por otra parte, para los proyectos considerados como no viables se nota una diferencia mayor; en este sentido cabe destacar los siguientes resultados: para P33 se obtuvo un error relativo aproximado de 47% (casi 1,40 unidades de diferencia), para P35 de casi 40% (aproximadamente 1,20 unidades) y para P37 de aproximadamente el 30% (con también 1,20 unidades de diferencia). En última instancia, y de acuerdo a lo que se ilustra en dicha figura, es importante destacar que el 76% de los proyectos tienen un error relativo menor al 10% confirmando así la precisión de la viabilidad global calculada por el modelo, en término de la poca diferencia existente entre los valores calculados y los indicados por los investigadores. En tal sentido, sólo existen variaciones mayores al 10% de error relativo para los siguientes proyectos P23, P26, P28, P30, P32, P33, P35, P36 y P37 (siendo que estos nueve proyectos constituyen el 24% del total).

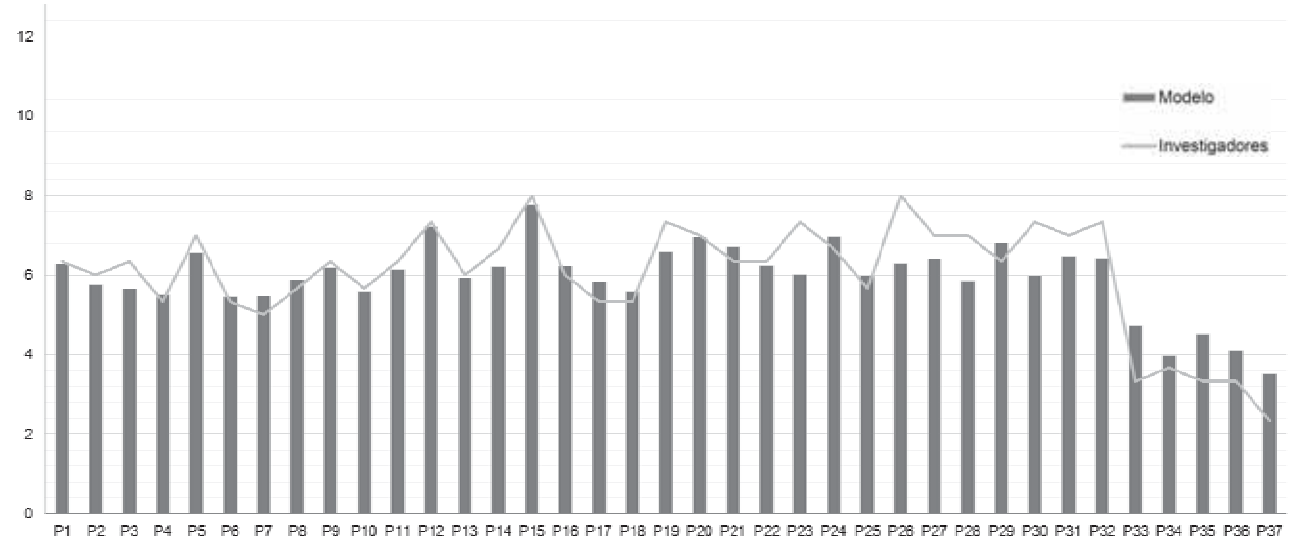


Fig. 5.8. Gráfico de comparación de los valores de la Viabilidad Global.

5.2.3. Prueba de Wilcoxon del Modelo para la Evaluación de la Viabilidad

Como último paso de la validación del modelo para la evaluación de la viabilidad, se aplica la prueba de rangos con signo de Wilcoxon [1945]. Esta prueba no paramétrica se utiliza para comprobar que no hay diferencia significativa entre la valoración realizada por los investigadores con los valores calculados por el modelo. Por consiguiente, a partir de los valores de la Tabla 5.1 tomados en forma de pares para cada una de las dimensiones de plausibilidad, adecuación, éxito y viabilidad global, se aplica esta prueba con el objetivo de comprobar que no hay diferencia entre los datos reales y los calculados en cada una de las dimensiones. Una descripción detallada del funcionamiento de esta prueba se encuentra disponible también en el Anexo F de este trabajo.

Asimismo, para cada una de las dimensiones se cumple con los requisitos necesarios para llevar a cabo cada prueba debido a que los proyectos poseen datos apareados que se seleccionaron aleatoriamente, y las diferencias entre los pares de datos (los proporcionados por los investigadores y los calculados por el modelo) tienen una distribución que es aproximadamente simétrica (tal como se puede observar en los gráficos que se acompañan en la prueba realizada para cada dimensión).

Para realizar cada prueba se aplica la siguiente hipótesis nula (H_0) y alternativa (H_1):

H_0: Los valores indicados por los investigadores y los calculados por el modelo son tales que la mediana de la población de las diferencias es igual a cero; es decir, no hay diferencias significativas entre lo indicado por los investigadores y lo generado por el modelo.

H_1: La mediana de la población de diferencias no es igual a cero; es decir, que existen diferencias significativas entre lo indicado por los investigadores y lo generado por el modelo.

El objetivo de cada prueba es probar si la hipótesis nula es válida o debe ser refutada, a tal efecto se utiliza un nivel de significancia de 0,01 (lo que equivale a un grado de confianza del 99%). Como se cuenta con 37 proyectos reales o pares ($n=37$) y el nivel de significancia seleccionado (α) es de 0,01 entonces la hipótesis nula (H_0) será rechazada si la menor suma de rangos (W) es menor o igual a 182 (valor crítico obtenido de la tabla estadística correspondiente). En caso contrario, no se rechaza y se considera como válida.

A continuación se presenta la aplicación de la prueba de Wilcoxon sobre los valores de cada dimensión del modelo propuesto:

- Prueba de la *Plausibilidad*:
  En la Figura 5.9 se presenta la dispersión de los rangos con signos para cada proyecto analizado. Con este gráfico es posible confirmar que la distribución de diferencias es simétrica para los datos considerados por lo que es posible aplicar la prueba de Wilcoxon. Los resultados de la prueba para esta dimensión se muestran en la Tabla 5.6.

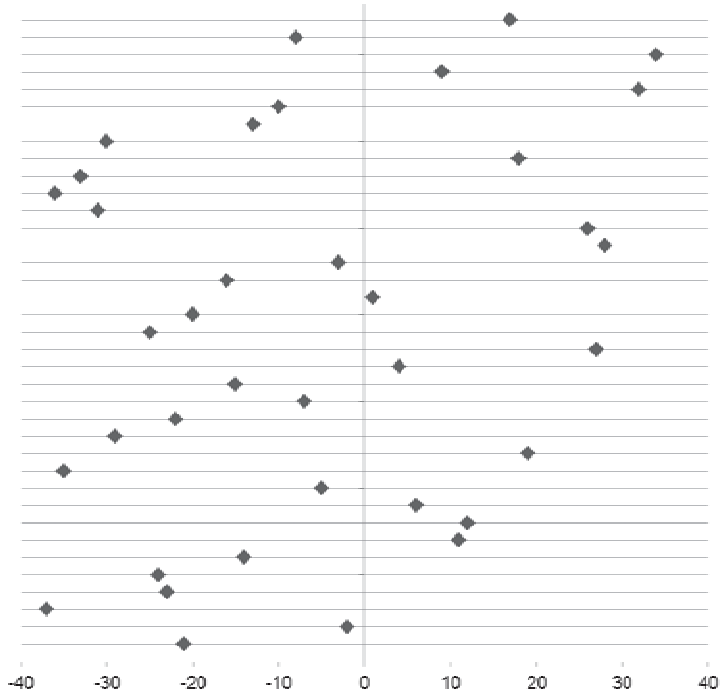


Fig. 5.9. Dispersión de los rangos con signo de la prueba para la dimensión Plausibilidad.

#	Valor calculado por el Modelo de Viabilidad	Valoración indicada por los Investigadores	Diferencias	Rango de las Diferencias	Rangos con signo	Rangos Positivos (W+)	Rangos Negativos (W-)
P1	7,20	8	- 0,80	21	- 21		*21*
P2	6,87	7	- 0,13	2	- 2		*2*
P3	5,90	8	- 2,10	37	- 37		*37*
P4	5,12	6	- 0,88	23	- 23		*23*
P5	5,12	6	- 0,88	24	- 24		*24*
P6	5,45	6	- 0,55	14	- 14		*14*
P7	5,45	5	0,45	11	11	*11*	
P8	6,45	6	0,45	12	12	*12*	
P9	7,20	7	0,20	6	6	*6*	
P10	5,85	6	- 0,15	5	- 5		*5*
P11	6,22	8	- 1,78	35	- 35		*35*
P12	7,67	7	0,67	19	19	*19*	
P13	5,93	7	- 1,07	29	- 29		*29*
P14	6,20	7	- 0,80	22	- 22		*22*
P15	8,72	9	- 0,28	7	- 7		*7*
P16	6,45	7	- 0,55	15	- 15		*15*
P17	6,14	6	0,14	4	4	*4*	
P18	6,00	5	1,00	27	27	*27*	
P19	7,01	8	- 0,99	25	- 25		*25*
P20	8,24	9	- 0,76	20	- 20		*20*
P21	8,05	8	0,05	1	1	*1*	
P22	6,45	7	- 0,55	16	- 16		*16*
P23	6,87	7	- 0,13	3	- 3		*3*
P24	8,05	7	1,05	28	28	*28*	
P25	6,00	5	1,00	26	26	*26*	
P26	6,55	8	- 1,45	31	- 31		*31*
P27	6,00	8	- 2,00	36	- 36		*36*
P28	6,39	8	- 1,61	33	- 33		*33*
P29	7,64	7	0,64	18	18	*18*	
P30	6,87	8	- 1,13	30	- 30		*30*
P31	6,52	7	- 0,48	13	- 13		*13*
P32	6,60	7	- 0,40	10	- 10		*10*
P33	4,49	3	1,49	32	32	*32*	
P34	4,36	4	0,36	9	9	*9*	
P35	4,66	3	1,66	34	34	*34*	
P36	4,66	5	- 0,34	8	- 8		*8*
P37	4,63	4	0,63	17	17	*17*	
Suma Total						**244**	**459**

Tabla 5.6. Resultados de prueba de Wilcoxon para la dimensión Plausibilidad.

En esta dimensión, para comprobar la hipótesis nula se utiliza la menor suma de rangos que corresponde a la de rangos positivos (W+); por consiguiente, la menor suma de rangos

considerada (W) es igual a 244. Dado que este valor es mayor que el valor crítico obtenido para el nivel de significancia seleccionado (182), no se rechaza la hipótesis nula (H_0). Esto significa que no hay diferencia significativa entre el valor calculado por el modelo para la dimensión plausibilidad y el indicado por los investigadores o, en otros términos, que el modelo evalúa el grado de plausibilidad del proyecto en forma similar a lo realizado por los expertos del dominio.

- Prueba de la *Adecuación*:

  En esta dimensión se vuelve a aplicar la prueba para la dimensión adecuación cuya dispersión simétrica se observa en la Figura 5.10 y sus resultados se muestran en la Tabla 5.7.

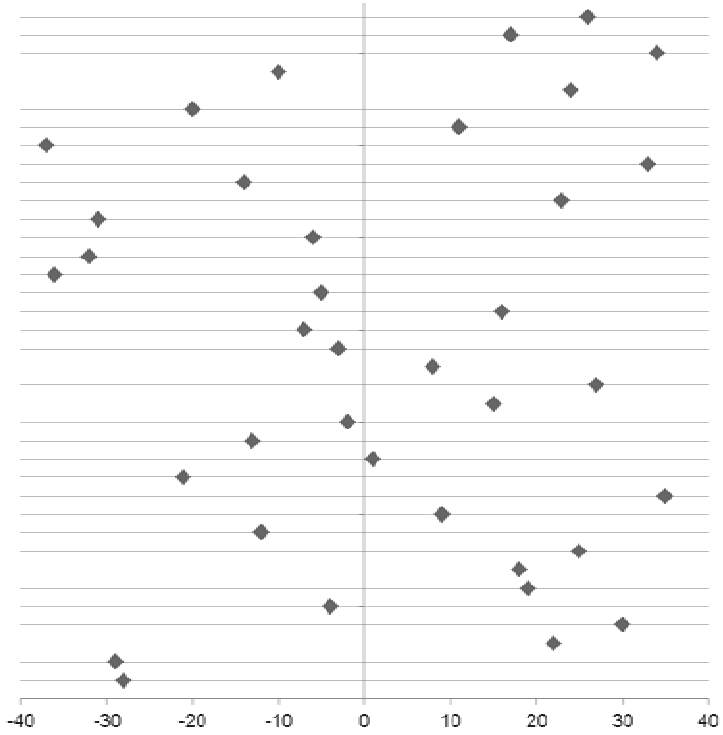


Fig. 5.10. Dispersión de los rangos con signo de la prueba para la dimensión Adecuación.

Para esta dimensión la suma de rangos negativos (W-) es la menor, por lo que se utiliza W = 310. Dado que este valor es mayor que 182, no se rechaza la hipótesis nula (H_0) y se puede decir con un nivel de significancia del 0,01 que no hay diferencia entre el valor calculado por el modelo para la dimensión adecuación y el suministrado por los investigadores.

#	Valor calculado por el Modelo de Viabilidad	Valoración indicada por los Investigadores	Diferencias	Rango de las Diferencias	Rangos con signo	Rangos Positivos (W+)	Rangos Negativos (W-)
P1	6,11	7	- 0,89	28	- 28		*28*
P2	5,07	6	- 0,93	29	- 29		*29*
P3	5,67	5	0,67	22	22	*22*	
P4	6,95	6	0,95	30	30	*30*	
P5	7,82	8	- 0,18	4	- 4		*4*
P6	5,61	5	0,61	19	19	*19*	
P7	5,56	5	0,56	18	18	*18*	
P8	5,80	5	0,80	25	25	*25*	
P9	5,61	6	- 0,39	12	- 12		*12*
P10	5,34	5	0,34	9	9	*9*	
P11	6,56	5	1,56	35	35	*35*	
P12	7,35	8	- 0,65	21	- 21		*21*
P13	5,09	5	0,09	1	1	*1*	
P14	6,59	7	- 0,41	13	- 13		*13*
P15	6,89	7	- 0,11	2	- 2		*2*
P16	6,43	6	0,43	15	15	*15*	
P17	5,83	5	0,83	27	27	*27*	
P18	5,31	5	0,31	8	8	*8*	
P19	6,89	7	- 0,11	3	- 3		*3*
P20	6,75	7	- 0,25	7	- 7		*7*
P21	6,45	6	0,45	16	16	*16*	
P22	5,81	6	- 0,19	5	- 5		*5*
P23	5,20	7	- 1,80	36	- 36		*36*
P24	6,76	8	- 1,24	32	- 32		*32*
P25	6,76	7	- 0,24	6	- 6		*6*
P26	7,00	8	- 1,00	31	- 31		*31*
P27	6,70	6	0,70	23	23	*23*	
P28	5,58	6	- 0,42	14	- 14		*14*
P29	6,27	5	1,27	33	33	*33*	
P30	5,90	8	- 2,10	37	- 37		*37*
P31	6,39	6	0,39	11	11	*11*	
P32	6,39	7	- 0,61	20	- 20		*20*
P33	4,77	4	0,77	24	24	*24*	
P34	4,62	5	- 0,38	10	- 10		*10*
P35	5,34	4	1,34	34	34	*34*	
P36	3,46	3	0,46	17	17	*17*	
P37	2,81	2	0,81	26	26	*26*	
Suma Total						**393**	**310**

Tabla 5.7. Resultados de prueba de Wilcoxon para la dimensión Adecuación.

- Prueba del *Éxito*:

  Los resultados de aplicar la prueba para la dimensión éxito se muestran en la Tabla 5.8 y la Figura 5.11.

#	Valor calculado por el Modelo de Viabilidad	Valoración indicada por los Investigadores	Diferencias	Rango de las Diferencias	Rangos con signo	Rangos Positivos (W+)	Rangos Negativos (W-)
P1	5,25	4	1,25	27	27	*27*	
P2	5,25	5	0,25	3	3	*3*	
P3	5,31	6	- 0,69	22	- 22		*22*
P4	4,12	4	0,12	1	1	*1*	
P5	6,81	7	- 0,19	2	- 2		*2*
P6	5,25	5	0,25	4	4	*4*	
P7	5,42	5	0,42	9	9	*9*	
P8	5,18	6	- 0,82	24	- 24		*24*
P9	5,57	6	- 0,43	11	- 11		*11*
P10	5,57	6	- 0,43	12	- 12		*12*
P11	5,42	6	- 0,58	18	- 18		*18*
P12	6,45	7	- 0,55	17	- 17		*17*
P13	7,05	6	1,05	26	26	*26*	
P14	5,69	6	- 0,31	7	- 7		*7*
P15	7,66	8	- 0,34	8	- 8		*8*
P16	5,64	5	0,64	20	20	*20*	
P17	5,42	5	0,42	10	10	*10*	
P18	5,42	6	- 0,58	19	- 19		*19*
P19	5,58	7	- 1,42	28	- 28		*28*
P20	5,52	5	0,52	14	14	*14*	
P21	5,25	5	0,25	5	5	*5*	
P22	6,54	6	0,54	15	15	*15*	
P23	5,96	8	- 2,04	34	- 34		*34*
P24	5,81	5	0,81	23	23	*23*	
P25	5,00	5	0,00	-	-	-	-
P26	5,01	8	- 2,99	36	- 36		*36*
P27	6,54	7	- 0,46	13	- 13		*13*
P28	5,47	7	- 1,53	30	- 30		*30*
P29	6,45	7	- 0,55	16	- 16		*16*
P30	4,97	6	- 1,03	25	- 25		*25*
P31	6,54	8	- 1,46	29	- 29		*29*
P32	6,20	8	- 1,80	31	- 31		*31*
P33	4,99	3	1,99	32	32	*32*	
P34	2,64	2	0,64	21	21	*21*	
P35	3,25	3	0,25	6	6	*6*	
P36	4,21	2	2,21	35	35	*35*	
P37	3,01	1	2,01	33	33	*33*	
Suma Total						**284**	**382**

Tabla 5.8. Resultados de prueba de Wilcoxon para la dimensión Éxito.

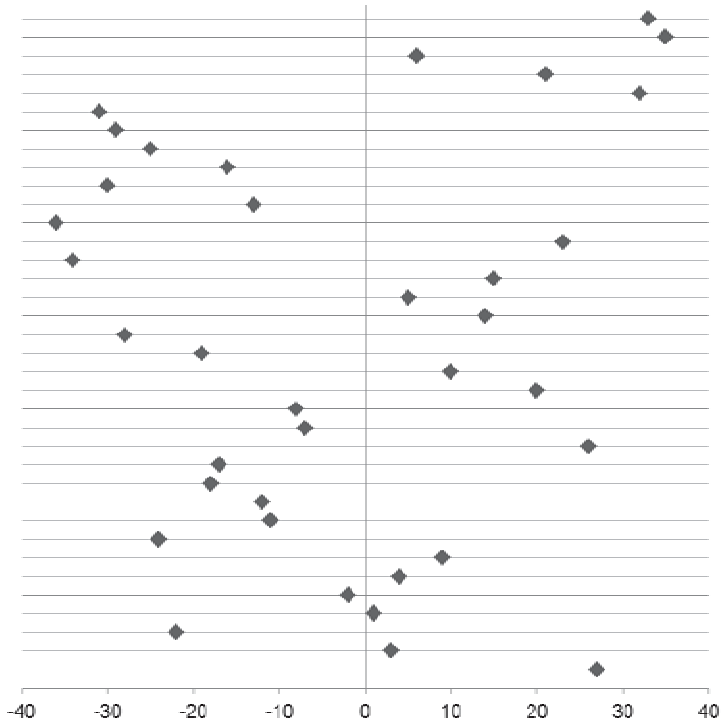


Fig. 5.11. Dispersión de los rangos con signo de la prueba para la dimensión Éxito.

Cabe destacar que en esta dimensión existe una la diferencia igual a cero entre el valor indicado por los investigadores y el calculado por el modelo para el proyecto P25. Por consiguiente, la cantidad de pares a ser considerados es de 36 y el valor crítico que se debe comparar con la menor suma de rangos es de 171 aplicando nuevamente un nivel de significancia del 0,01.

Como W = W+ = 284 y este valor también es mayor a 171, no se rechaza H_0; se puede decir que no hay diferencia entre el valor calculado por el modelo para esta dimensión y el asignado por los investigadores.

- Prueba de la *Viabilidad Global del Proyecto*:

  Finalmente, se comparan los valores indicados con los calculados para la viabilidad global del proyecto. Los resultados de esta prueba se muestran en la Tabla 5.9 y la Figura 5.12.

  Para esta dimensión (que se calcula en base las anteriores) la distribución de valores es mucho más simétrica que con las dimensiones anteriores. Esto se nota tanto por la cercanía entre los valores W+ y W- como por la dispersión de puntos en el gráfico de la Figura 5.12.

#	Valor calculado por el Modelo de Viabilidad	Valoración indicada por los Investigadores	Diferencias	Rango de las Diferencias	Rangos con signo	Rangos Positivos (W+)	Rangos Negativos (W-)
P1	6,27	6,33	-0,06	2	-2		*2*
P2	5,77	6,00	-0,23	13	-13		*13*
P3	5,65	6,33	-0,68	27	-27		*27*
P4	5,51	5,33	0,18	9	9	*9*	*0*
P5	6,56	7,00	- 0,44	20	- 20		*20*
P6	5,45	5,33	0,12	7	7	*7*	
P7	5,48	5,00	0,48	22	22	*22*	
P8	5,87	5,67	0,20	11	11	*11*	
P9	6,18	6,33	- 0,15	8	- 8		*8*
P10	5,59	5,67	- 0,08	4	- 4		*4*
P11	6,14	6,33	- 0,19	10	- 10		*10*
P12	7,22	7,33	- 0,11	6	- 6		*6*
P13	5,93	6,00	- 0,07	3	- 3		*3*
P14	6,20	6,67	- 0,47	21	- 21		*21*
P15	7,77	8,00	- 0,23	14	- 14		*14*
P16	6,22	6,00	0,22	12	12	*12*	
P17	5,83	5,33	0,50	24	24	*24*	
P18	5,59	5,33	0,26	15	15	*15*	
P19	6,58	7,33	- 0,75	28	- 28		*28*
P20	6,96	7,00	- 0,04	1	- 1		*1*
P21	6,70	6,33	0,37	19	19	*19*	
P22	6,24	6,33	- 0,09	5	- 5		*5*
P23	6,01	7,33	- 1,32	34	- 34		*34*
P24	6,97	6,67	0,30	16	16	*16*	
P25	6,00	5,67	0,33	18	18	*18*	
P26	6,29	8,00	- 1,71	37	- 37		*37*
P27	6,40	7,00	- 0,60	26	- 26		*26*
P28	5,85	7,00	- 1,15	31	- 31		*31*
P29	6,82	6,33	0,48	23	23	*23*	
P30	6,00	7,33	- 1,34	35	- 35		*35*
P31	6,48	7,00	- 0,52	25	- 25		*25*
P32	6,42	7,33	- 0,92	30	- 30		*30*
P33	4,73	3,33	1,39	36	36	*36*	
P34	3,99	3,67	0,32	17	17	*17*	
P35	4,52	3,33	1,19	32	32	*32*	
P36	4,10	3,33	0,77	29	29	*29*	
P37	3,52	2,33	1,19	33	33	*33*	
Suma Total						**323**	**380**

Tabla 5.9. Resultados de prueba de Wilcoxon para la Viabilidad Global del Proyecto.

Al aplicar la prueba, ya que W = W+ = 323 > 182, no se rechaza $H_{0;}$ se puede decir con un grado de confianza del 99% que no hay diferencia entre el valor calculado por el modelo para la Viabilidad Global del Proyecto y el indicado por los investigadores.

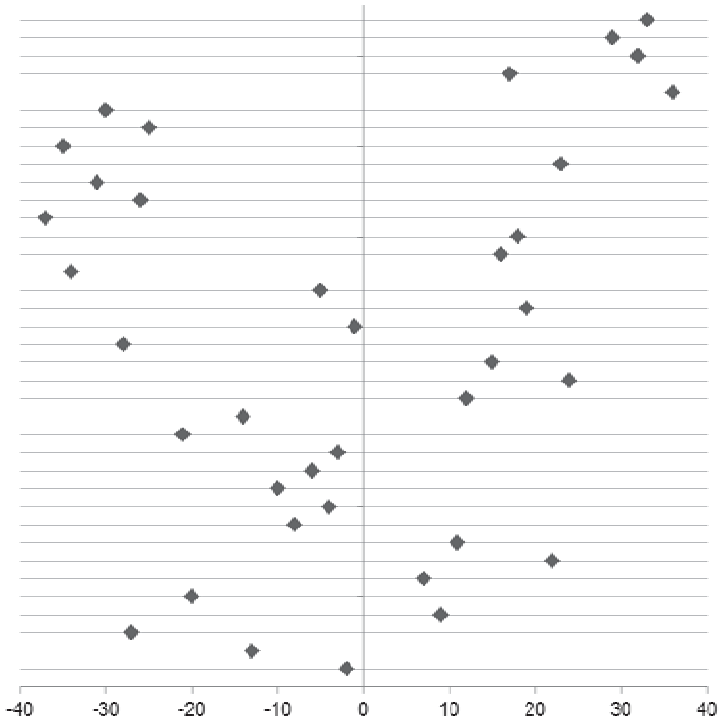


Fig. 5.12. Dispersión de los rangos con signo de la prueba para la Viabilidad Global del Proyecto.

5.2.4. Conclusiones de la Validación del Modelo para la Evaluación de la Viabilidad

A partir de los análisis estadísticos realizados, se observa que el modelo posee un comportamiento muy similar a lo definido por los investigadores para los proyectos considerados. Aunque el modelo suele ser un poco más conservador para determinar valores bajos, el comportamiento general es preciso. En todo caso, se puede decir que los investigadores fueron más críticos, teniendo un punto de vista más pesimista al momento de asignar los valores de las dimensiones.

Esto hecho se confirma posteriormente en virtud de los resultados obtenidos a partir de las pruebas de Wilcoxon realizadas. Con un grado de confianza del 99%, para todas las dimensiones no se ha detectado diferencia significativa entre el valor calculado por el modelo y el asignado en la valoración realizada por los investigadores. La mayor simetría se detecta para la viabilidad global del proyecto, la cual se calcula en base a las otras tres dimensiones. A partir de este hecho significa que no obstante para algunos casos el resultado de la valoración de una dimensión no sea exactamente igual a lo asignado por los investigadores, al calcular la viabilidad global del proyecto el modelo compensa las diferencias para obtener un resultado más aproximado al valor real.

En función de lo expuesto, es razonable inferir que el Modelo para la Evaluación de la Viabilidad propuesto es confiable a los efectos de ser utilizado en proyectos de Explotación de Información dentro del ámbito de las PyMEs.

5.3. VALIDACIÓN DEL MODELO PARA LA ESTIMACIÓN DE ESFUERZO DE PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

En esta sección se presenta la validación del segundo modelo propuesto que puede ser utilizado para estimar el esfuerzo requerido para desarrollar un proyecto de Explotación de Información dentro del ámbito de una PyME. Esta sección se encuentra estructurada de la siguiente forma: los datos utilizados para llevar a cabo esta validación (sección 5.3.1), la realización de un análisis estadístico sobre dichos datos (sección 5.3.2), la aplicación de la prueba de Wilcoxon (sección 5.3.3); y la presentación de las conclusiones preliminares acerca del modelo (5.3.4).

5.3.1. Datos utilizados en la Validación del Modelo para la Estimación de Esfuerzo

Tal como se llevó a cabo para el modelo de viabilidad, la validación de éste consiste en comparar del esfuerzo estimado al aplicar el modelo con el esfuerzo real que fue requerido para desarrollar los proyectos en forma completa. En este sentido, para esta validación se toman sólo los proyectos suministrados por los investigadores que fueron finalizados satisfactoriamente (P1 a P32); dejando de lado los proyectos que han sido cancelados, dado que el esfuerzo real es desconocido para estos últimos proyectos. A cada uno de estos treinta y dos proyectos que finalizaron satisfactoriamente se les aplica los dos métodos de estimación incluidos en el modelo, tal como se describe en el anexo D de este trabajo de tesis.

Asimismo y a modo de resumen, se presentan en la Tabla 5.10 los valores obtenidos de cada uno de los respectivos proyectos. En dicha tabla se incluye tanto el esfuerzo real que fue requerido para realizar el proyecto en forma completa (ER), como así también los esfuerzos estimados por el modelo propuesto mediante la fórmula lineal (PEM_L) y el método empírico (PEM_E). Por otra parte, para cada esfuerzo estimado por el modelo se calcula el error absoluto (la diferencia entre el esfuerzo real y el calculado) y el esfuerzo relativo (calculado como el error dividido por el esfuerzo real); recordando que todos estos valores se expresan en meses/hombre.

Al observar la Tabla 5.10, es importante observar que la fórmula lineal genera un error mucho menor al método empírico. Mientras que el promedio del error absoluto para el primer método es de 0,89 meses/hombre (con un desvío del error de ± 1,53 meses/hombre), para el segundo es de aproximadamente 1,52 meses/hombre (con un desvío de aproximadamente ± 2,21 meses/hombre). Un análisis más detallado de estos valores se lleva a cabo en las secciones siguientes.

#	Esfuerzo Real ER	Fórmula Lineal			Método Empírico		
		Esfuerzo calculado PEM_L	Error $ER - PEM_L$	Error Relativo $\frac{ER - PEM_L}{ER}$	Esfuerzo calculado PEM_E	Error $ER - PEM_E$	Error Relativo $\frac{ER - PEM_E}{ER}$
P1	*2,41*	*2,58*	- 0,17	- 7,1%	*3,57*	- 1,16	- 48,1%
P2	*7,00*	*6,00*	1,00	14,3%	*7,23*	- 0,23	- 3,3%
P3	*1,64*	*1,48*	0,16	9,8%	*3,58*	- 1,94	- 118,3%
P4	*3,65*	*1,68*	1,97	54,0%	*3,57*	0,08	2,2%
P5	*9,35*	*9,80*	- 0,45	- 4,8%	*7,58*	1,77	18,9%
P6	*11,63*	*5,10*	6,53	56,1%	*6,07*	5,56	47,8%
P7	*6,73*	*3,78*	2,95	43,8%	*5,91*	0,82	12,2%
P8	*5,40*	*4,88*	0,52	9,6%	*3,06*	2,34	43,3%
P9	*8,38*	*8,70*	- 0,32	- 3,8%	*7,66*	0,72	8,5%
P10	*1,56*	*1,08*	0,48	30,8%	*1,50*	0,06	4,1%
P11	*9,70*	*9,60*	0,10	1,0%	*12,64*	- 2,94	- 30,3%
P12	*5,24*	*5,80*	- 0,56	- 10,7%	*5,63*	- 0,39	- 7,4%
P13	*5,00*	*4,58*	0,42	8,4%	*3,17*	1,84	36,7%
P14	*8,97*	*9,18*	- 0,21	- 2,3%	*5,91*	3,06	34,1%
P15	*2,81*	*3,48*	- 0,67	- 23,8%	*1,11*	1,70	60,4%
P16	*11,80*	*12,00*	- 0,20	- 1,7%	*7,58*	4,22	35,7%
P17	*2,79*	*2,28*	0,51	18,3%	*8,44*	- 5,65	- 202,5%
P18	*3,88*	*3,58*	0,30	7,7%	*3,57*	0,31	8,0%
P19	*5,70*	*6,30*	- 0,60	- 10,5%	*10,11*	- 4,41	- 77,4%
P20	*8,54*	*9,18*	- 0,64	- 7,5%	*8,44*	0,10	1,2%
P21	*10,61*	*11,50*	- 0,89	- 8,4%	*7,33*	3,28	30,9%
P22	*6,88*	*6,40*	0,48	7,0%	*6,71*	0,17	2,5%
P23	*11,20*	*9,70*	1,50	13,4%	*10,11*	1,09	9,7%
P24	*9,70*	*12,70*	- 3,00	- 30,9%	*10,93*	- 1,23	- 12,7%
P25	*7,30*	*8,38*	- 1,08	- 14,8%	*6,51*	0,80	10,9%
P26	*5,31*	*5,10*	0,21	4,0%	*5,63*	- 0,32	- 6,0%
P27	*6,10*	*6,70*	- 0,60	- 9,8%	*5,63*	0,47	7,7%
P28	*10,00*	*9,60*	0,40	4,0%	*9,39*	0,61	6,1%
P29	*6,43*	*7,12*	- 0,69	- 10,7%	*5,91*	0,52	8,1%
P30	*9,80*	*10,20*	- 0,40	- 4,1%	*10,11*	- 0,31	- 3,2%
P31	*1,50*	*1,68*	- 0,18	- 12,0%	*1,11*	0,39	25,8%
P32	*3,78*	*3,42*	0,36	9,5%	*4,04*	- 0,26	- 6,8%

Tabla 5.10. Datos de los proyectos usados en la validación del modelo de estimación de esfuerzo (en meses/hombre).

Por otra parte, y en forma complementaria, en el anexo E se indican los resultados de aplicar el modelo de estimación DMCoMo [Marbán *et al.*, 2008] sobre los mismos proyectos. Aplicando DMCoMo, los esfuerzos estimados son siempre superiores a los reales generando de esta manera una sobreestimación; siendo en este caso el error más pequeño de 24 meses/hombre para el proyecto P16 con la fórmula MM23. Por consiguiente, los resultados obtenidos por los métodos del modelo propuesto son mucho más precisos que los de DMCoMo en este tipo de proyectos.

5.3.2. Análisis Estadístico del Modelo para la Estimación de Esfuerzo

Teniendo en cuenta los datos indicados en la Tabla 5.10, en esta sección se lleva a cabo un análisis estadístico del modelo de estimación de esfuerzo. Dicho análisis consiste en la comparación del esfuerzo real que se requirió para desarrollar el proyecto de forma completa con el valor calculado por cada método del modelo propuesto. En tal sentido, este análisis estadístico se lleva a cabo considerando a cada método de forma independiente.

- Análisis de la *Fórmula Lineal de estimación*:
  En este análisis e comparan los resultados estadísticos entre el esfuerzo real de cada proyecto y el calculado por la fórmula obtenida a partir del método de regresión lineal (PEM_L). Los resultados obtenidos se indican en la Tabla 5.11 y, para facilitar la interpretación, también se reflejan en el gráfico Boxplot de la Figura 5.13.

Parámetro Estadístico	**Esfuerzo Real**	**PEM_L**
Valor Mínimo	*1,50*	*1,08*
Valor Máximo	*11,80*	*12,70*
Valor Promedio	*6,59*	*6,36*
Valor de Varianza	*9,89*	*11,29*

Tabla 5.11. Resultados estadísticos para la Fórmula Lineal de estimación.

Como se puede observar en la Tabla 5.11, la fórmula lineal tiende a estimar esfuerzos muy similares a los reales; aunque con valores un tanto más dispersos que los encontrados en la realidad. Se nota que la fórmula propuesta tiende a generar estimaciones un poco inferiores y superiores a las reales, teniendo una diferencia de aproximadamente 0,42 meses/hombre entre los valores mínimos y 0,90 meses/hombre con los máximos. En líneas generales el error

promedio es menor a 0,90 meses/hombre; y por lo tanto se puede considerar que la fórmula lineal de estimación posee un aceptable nivel de precisión.

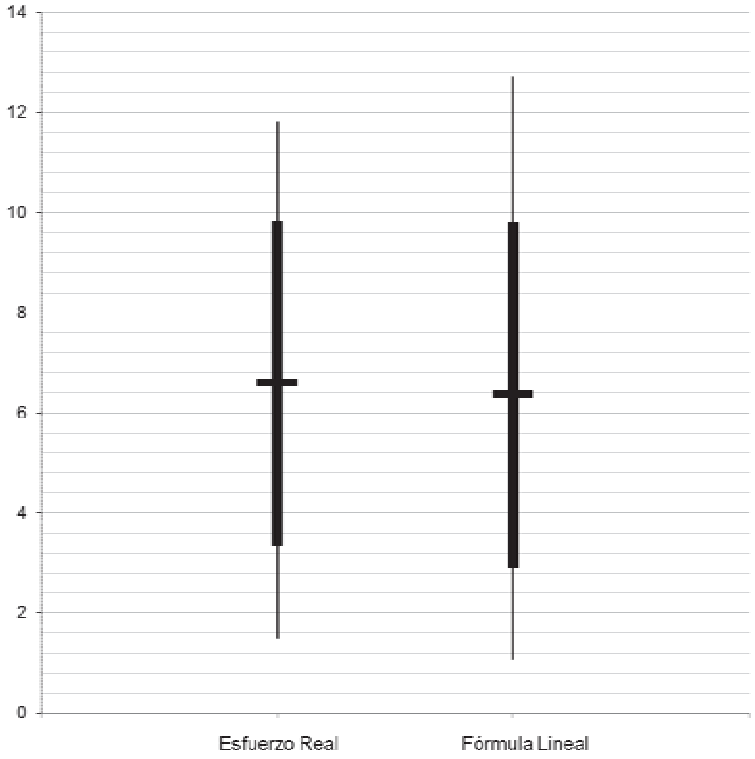


Fig. 5.13. Gráfico boxplot de la Fórmula Lineal de estimación.

Con el objetivo de presentar un análisis más detallado de esta fórmula, se genera el gráfico de la Figura 5.14 donde se ilustra, para cada proyecto, la diferencia entre el esfuerzo real (representado en forma de línea) con el esfuerzo calculado por la fórmula lineal (representado en forma de barra).

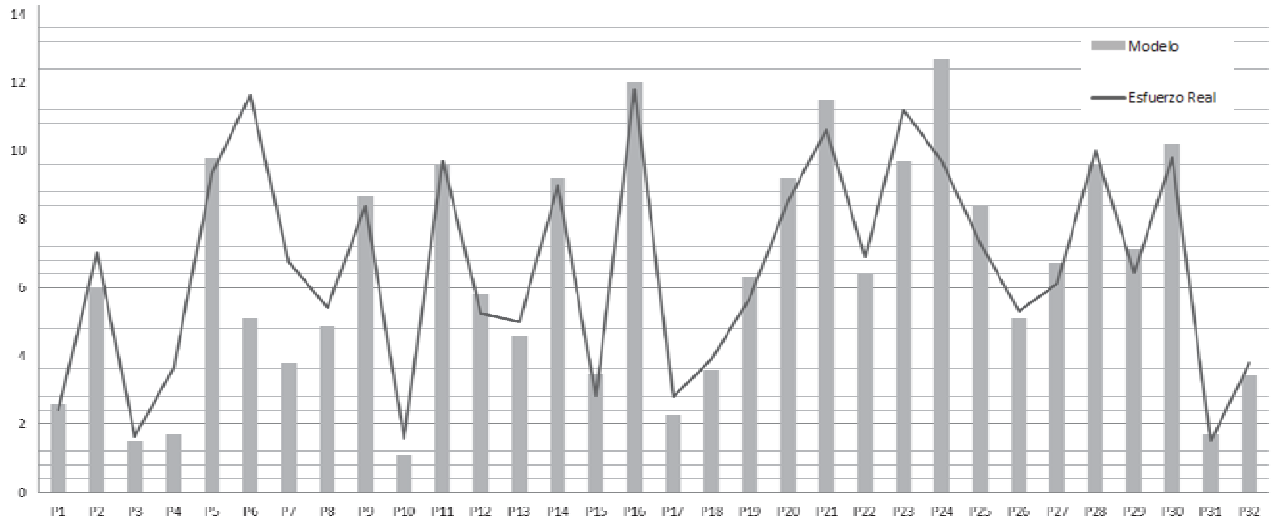


Fig. 5.14. Gráfico de comparación de los valores de la Fórmula Lineal de estimación.

En la Figura 5.14 se puede confirmar que esta fórmula genera resultados muy precisos. En tal sentido, para el 56% de los proyectos analizados (dieciocho proyectos de los treinta y dos proyectos), el error relativo es menor al 10%; mientras que para el 84% (veintisiete de los proyectos) el error relativo es menor al 25% del esfuerzo real.

El mayor error obtenido es de 6,53 meses/hombre y corresponde al proyecto P6. De acuerdo con la información suministrado por los administradores del proyecto, el mismo fue finalizado con éxito, aunque requirió un esfuerzo adicional debido a circunstancias inesperadas durante la preparación de los datos y la interpretación de los resultados. Entre los motivos de estas circunstancias se puede destacar la falta de experiencia previa del equipo de trabajo en el ámbito de la medicina. Por otra parte, otros proyectos que presentaron errores en su estimación (aunque mucho menores al de P6) son: P24 con una sobreestimación de 3 meses/hombre, P7 con una subestimación de 2,95 meses/hombre y P4 con 1,97 meses/hombre, siendo el error absoluto del resto menor o igual a 1,50 meses/hombre.

- Análisis del *Método Empírico de estimación*:

  De igual manera que para el caso de la fórmula anterior, en la Tabla 5.12 y en forma gráfica en la Figura 5.15 ilustran los resultados de la comparación entre el esfuerzo real y el calculado por el método empírico propuesto (PEM_E).

Parámetro Estadístico	**Esfuerzo Real**	**PEM_E**
Valor Mínimo	*1,50*	*1,11*
Valor Máximo	*11,80*	*12,64*
Valor Promedio	*6,59*	*6,24*
Valor de Varianza	*9,89*	*8,55*

Tabla 5.12. Resultados estadísticos para el Método Empírico de estimación.

Para el método empírico se puede observar que los esfuerzos estimados tienden a ser un tanto inferiores a los reales. La diferencia entre los valores mínimos y máximos es relativamente pequeña (de 0,42 meses/hombre entre los valores mínimos y 0,84 meses/hombre con los máximos). Con este método, el error promedio es un poco mayor que para el caso de la fórmula lineal (aproximadamente de 1,50 meses/hombre).

Asimismo, nuevamente se presenta el gráfico detallado para cada proyecto de la Figura 5.16. A partir del gráfico correspondiente a esta figura, se infiere que la aplicación del método empírico origina estimaciones con mayor error en relación a las obtenidas con la fórmula lineal. En este

sentido, sólo quince de los proyectos (P2, P4, P9, P10, P12, P18, P20, P22, P23, P26, P27, P28, P29, P30 y P32) poseen un error relativo menor al 10%. No obstante, la mayoría de ellos (veintitrés) poseen un error relativo menor al 35%. Es importante destacar que el 88% de los proyectos (veintiocho) poseen un error menor al 50%.

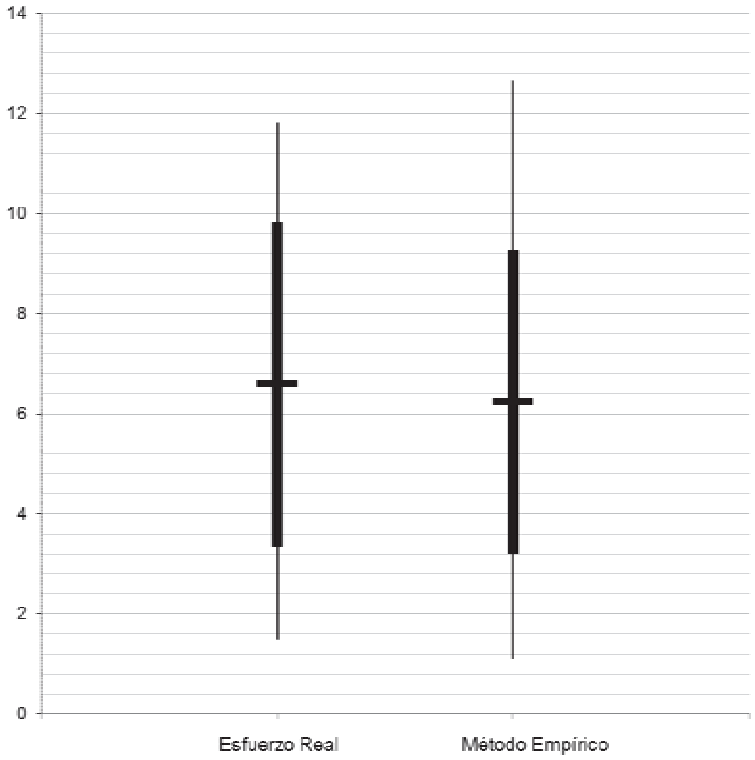


Fig. 5.15. Gráfico boxplot del Método Empírico de estimación.

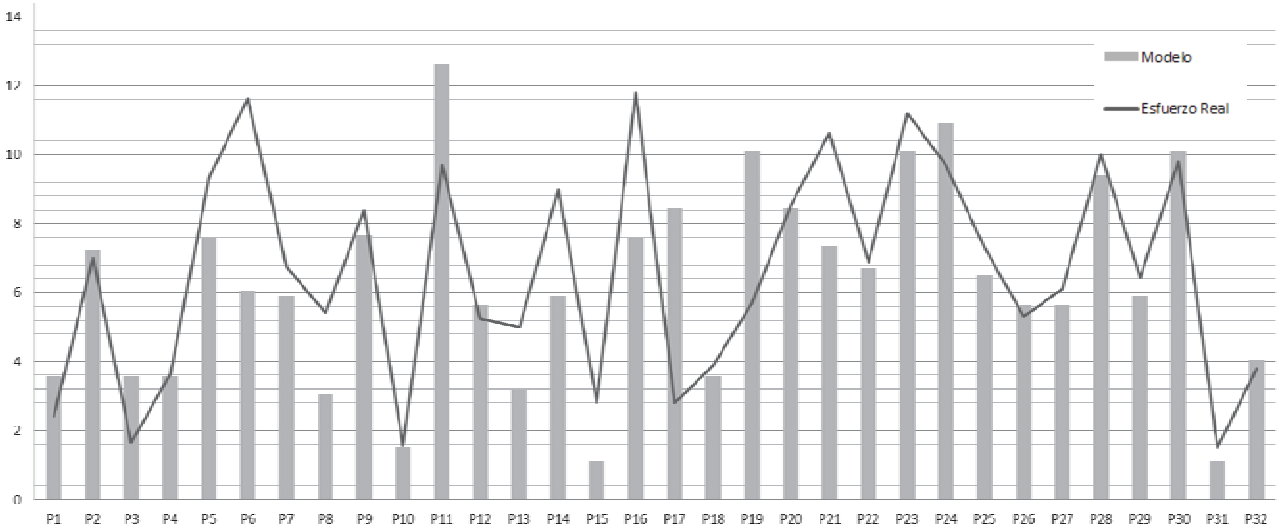


Fig. 5.16. Gráfico de comparación del Método Empírico de estimación.

Para el método empírico, el mayor error que se identifica para este método corresponde a P17 con una sobreestimación de 5,65 meses/hombre (el valor obtenido por el método empírico

supera en ese valor al esfuerzo real), destacándose también P6 con 5,56 meses/hombre, P19 con 4,41 y P16 con 4,22 meses/hombre. En el resto de los proyectos, el error absoluto es menor a 3,50 meses/hombre.

5.3.3. Prueba de Wilcoxon del Modelo para la Estimación de Esfuerzo

Para finalizar la validación del método de estimación de esfuerzo, se aplica nuevamente la prueba de rangos con signo de Wilcoxon sobre cada par de valores de la Tabla 5.10; tomando la fórmula lineal y el método de estimación empírico de manera independiente. Se recuerda que en el Anexo F de este trabajo de tesis se encuentra una descripción detallada del funcionamiento de esta prueba.
Para el modelo de estimación también las diferencias de los pares de datos tienen una distribución que es aproximadamente simétrica, cumpliendo así con el requerimiento solicitado por la prueba.
Para llevar a cabo esta nueva prueba se aplican la hipótesis nula (H_0) y la alternativa (H_1):

H_0: El esfuerzo real y el calculado por el modelo son tales que la mediana de la población de las diferencias es igual a cero; es decir, no hay diferencias significativas entre el esfuerzo requerido realmente y el estimado por el modelo.

H_1: La mediana de la población de diferencias no es igual a cero; es decir, que existen diferencias significativas entre el esfuerzo real requerido y el estimado por el modelo.

Para analizar dichas hipótesis se aplica nuevamente un nivel de significancia de 0,01 (lo que equivale a un grado de confianza del 99%). Como en este caso la cantidad de pares o proyectos reales es igual a 32, el valor crítico correspondiente de la tabla estadística para aceptar o rechazar la hipótesis nula es igual a 128.

A continuación se presenta la aplicación de la prueba para cada método:

- Prueba de la *Fórmula Lineal de estimación*:

  En primer término, se compara el esfuerzo real con el calculado por la fórmula lineal del modelo de estimación de esfuerzo. Los resultados de la aplicación de esta prueba se muestran en la Tabla 5.13 y la dispersión de los rangos con signos se representa en la Figura 5.17.

  Para la fórmula lineal, el mínimo valor de la suma de rangos (W) es igual a 262, que corresponde a la suma de los rangos negativos (W-). Dado que 262 es mayor a 128 (el valor crítico), no se rechaza la hipótesis nula (H_0) y se puede afirmar que no hay diferencias significativas entre el esfuerzo real y el calculado por la fórmula lineal. En otras palabras, la

fórmula lineal posee un comportamiento similar al esfuerzo real requerido para desarrollar el proyecto en forma completa.

#	Esfuerzo calculado por la Fórmula Lineal	Esfuerzo Real	Diferencias	Rango de Diferencias	Rangos con signo	Rangos Positivos (W+)	Rangos Negativos (W-)
P1	2,58	2,41	0,17	3	3	*3*	
P2	6,00	7,00	- 1,00	26	- 26		*26*
P3	1,48	1,64	- 0,16	2	- 2		*2*
P4	1,68	3,65	- 1,97	29	- 29		*29*
P5	9,80	9,35	0,45	14	14	*14*	
P6	5,10	11,63	- 6,53	32	- 32		*32*
P7	3,78	6,73	- 2,95	30	- 30		*30*
P8	4,88	5,40	- 0,52	18	- 18		*18*
P9	8,70	8,38	0,32	9	9	*9*	
P10	1,08	1,56	- 0,48	16	- 16		*16*
P11	9,60	9,70	- 0,10	1	- 1		*1*
P12	5,80	5,24	0,56	19	19	*19*	
P13	4,58	5,00	- 0,42	13	- 13		*13*
P14	9,18	8,97	0,21	7	7	*7*	
P15	3,48	2,81	0,67	23	23	*23*	
P16	12,00	11,80	0,20	5	5	*5*	
P17	2,28	2,79	- 0,51	17	- 17		*17*
P18	3,58	3,88	- 0,30	8	- 8		*8*
P19	6,30	5,70	0,60	21	21	*21*	
P20	9,18	8,54	0,64	22	22	*22*	
P21	11,50	10,61	0,89	25	25	*25*	
P22	6,40	6,88	- 0,48	15	- 15		*15*
P23	9,70	11,20	- 1,50	28	- 28		*28*
P24	12,70	9,70	3,00	31	31	*31*	
P25	8,38	7,30	1,08	27	27	*27*	
P26	5,10	5,31	- 0,21	6	- 6		*6*
P27	6,70	6,10	0,60	20	20	*20*	
P28	9,60	10,00	- 0,40	11	- 11		*11*
P29	7,12	6,43	0,69	24	24	*24*	
P30	10,20	9,80	0,40	12	12	*12*	
P31	1,68	1,50	0,18	4	4	*4*	
P32	3,42	3,78	- 0,36	10	- 10		*10*
Suma Total						**266**	**262**

Tabla 5.13. Resultados de prueba de Wilcoxon para la Fórmula Lineal de estimación.

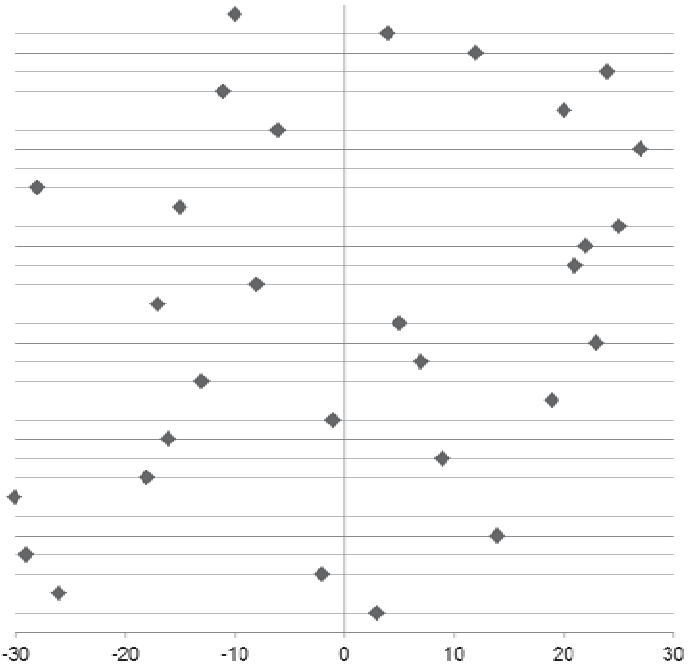


Fig. 5.17. Dispersión de los rangos con signo de la prueba para la Fórmula Lineal de estimación.

- Prueba del *Método Empírico de estimación*:

  En segundo término, se vuelve a aplicar la prueba con las estimaciones realizadas por el método empírico de estimación tal como se muestra en la Tabla 5.14. La dispersión correspondiente se indica en la Figura 5.18.

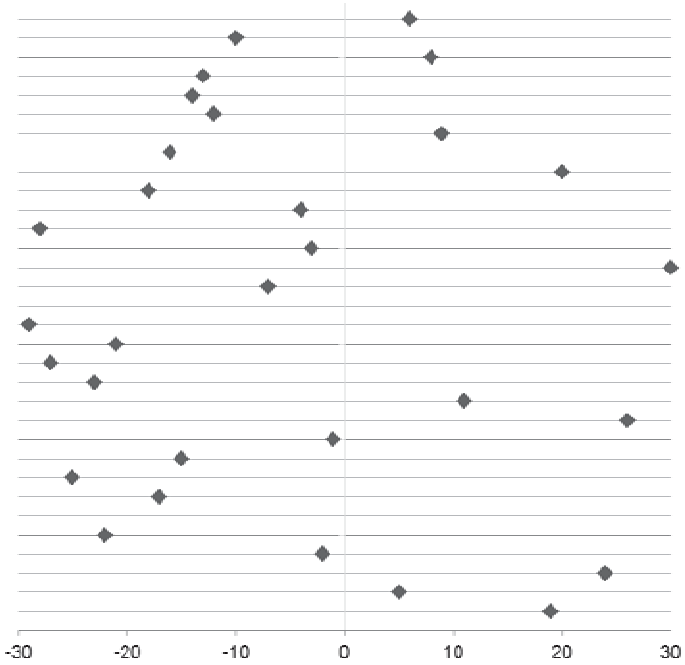


Fig. 5.18. Dispersión de los rangos con signo de la prueba para el Método Empírico de estimación.

#	Esfuerzo calculado por el Método Empírico	Esfuerzo Real	Diferencias	Rango de Diferencias	Rangos con signo	Rangos Positivos (W+)	Rangos Negativos (W-)
P1	3,57	2,41	1,16	19	19	*19*	
P2	7,23	7,00	0,23	5	5	*5*	
P3	3,58	1,64	1,94	24	24	*24*	
P4	3,57	3,65	-0,08	2	-2		*2*
P5	7,58	9,35	-1,77	22	-22		*22*
P6	6,07	11,63	-5,56	31	-31		*31*
P7	5,91	6,73	-0,82	17	-17		*17*
P8	3,06	5,40	-2,34	25	-25		*25*
P9	7,66	8,38	-0,72	15	-15		*15*
P10	1,50	1,56	-0,06	1	-1		*1*
P11	12,64	9,70	2,94	26	26	*26*	
P12	5,63	5,24	0,39	11	11	*11*	
P13	3,17	5,00	-1,84	23	-23		*23*
P14	5,91	8,97	-3,06	27	-27		*27*
P15	1,11	2,81	-1,70	21	-21		*21*
P16	7,58	11,80	-4,22	29	-29		*29*
P17	8,44	2,79	5,65	32	32	*32*	
P18	3,57	3,88	-0,31	7	-7		*7*
P19	10,11	5,70	4,41	30	30	*30*	
P20	8,44	8,54	-0,10	3	-3		*3*
P21	7,33	10,61	-3,28	28	-28		*28*
P22	6,71	6,88	-0,17	4	-4		*4*
P23	10,11	11,20	-1,09	18	-18		*18*
P24	10,93	9,70	1,23	20	20	*20*	
P25	6,51	7,30	-0,80	16	-16		*16*
P26	5,63	5,31	0,32	9	9	*9*	
P27	5,63	6,10	-0,47	12	-12		*12*
P28	9,39	10,00	-0,61	14	-14		*14*
P29	5,91	6,43	-0,52	13	-13		*13*
P30	10,11	9,80	0,31	8	8	*8*	
P31	1,11	1,50	-0,39	10	-10		*10*
P32	4,04	3,78	0,26	6	6	*6*	
Suma Total						**190**	**338**

Tabla 5.14. Resultados de prueba de Wilcoxon para el Método Empírico de estimación.

Dado que en este caso W = W+ = 190 y este valor también es mayor a 128, no se rechaza H_0 con un nivel de significancia del 0,01. Por consiguiente, se puede indicar que no hay diferencia significativa entre el esfuerzo calculado por el método empírico y el esfuerzo real empleado para realizar el proyecto.

5.3.4. Conclusiones de la Validación del Modelo para la Estimación de Esfuerzo

Con los resultados de los análisis presentados en las secciones anteriores, se observa que en líneas generales el modelo de Estimación de Esfuerzo posee un comportamiento similar al de los proyectos reales considerados. Se destaca la fórmula lineal de estimación por generar resultados muchos más precisos (su error promedio es inferior al mes/hombre); mientras en el caso del método empírico, el error es un poco mayor (alcanzando aproximadamente los 1,50 meses/hombre) y tendiendo a generar estimaciones de esfuerzo menores a las reales.

De todas formas, al comparar los resultados de ambos métodos con los del modelo DMCoMo (cuyos resultados de su aplicación a los mismos proyectos se incluye en el anexo E de este trabajo), se puede afirmar que los métodos propuestos son más precisos para estimar proyectos pequeños y medianos, debido a que DMCoMo se encuentra orientado para proyectos grandes.

De forma alternativa, mediante la prueba de Wilcoxon se ha confirmado que con un grado de confianza del 99% no hay diferencia significativa entre el esfuerzo calculado por los dos métodos del modelo propuesto y el esfuerzo real requerido para realizar el proyecto en forma completa. En tal sentido, se destaca una mayor simetría para la fórmula lineal de estimación, siendo la simetría del método empírico algo menor.

Por consiguiente, es razonable inferir que el Modelo para la Estimación de Esfuerzo propuesto proporciona un grado de confianza satisfactorio a los efectos de ser utilizado en proyectos de Explotación de Información dentro del ámbito de las PyMEs.

6. CONCLUSIONES

En este capítulo se describen las aportaciones de esta tesis doctoral (sección 6.1), y se destacan las futuras líneas de investigación que se consideran de interés en base a los problemas abiertos identificados (sección 6.2).

6.1. APORTACIONES DE LA TESIS

Este trabajo de tesis formula aportaciones al cuerpo de conocimientos de la Explotación de Información focalizados en la generación de información que permita asistir al proceso de gestión de estos proyectos desde su comienzo. En este sentido, se han tenido en cuenta las características generales de los proyectos de Explotación de Información, así como también sus particularidades, en aquellos desarrollos que tienen lugar en el ámbito de las Pequeñas y Medianas Empresas (PyMEs).

Teniendo en cuenta las preguntas de investigación planteadas en la sección 3.4 del presente trabajo de tesis, cabe citar las siguientes aportaciones:

I. Se ha determinado que es posible proponer un Modelo para la Evaluación de la Viabilidad de proyectos de Explotación de Información dentro del ámbito de las PyMEs, el cual contiene:

   a. La definición de trece condiciones (o preguntas) asociadas al proyecto que deben ser identificadas por el Ingeniero de Explotación de Información; dichas condiciones presentan las siguientes particularidades:

      - se clasifican en tres dimensiones teniendo en cuenta parte el aspecto del proyecto que evalúan: si es posible realizarlo (*Plausibilidad*), si la solución es la correcta (*Adecuación*), y si es posible cumplir con todas las metas y expectativas del mismo (*Éxito*).
      - permiten determinar las características asociadas a los datos disponibles, al problema de negocio que se desea resolver, al equipo de trabajo y al proyecto.
      - cabe destacar que estas características señaladas pueden conocerse al comienzo de un proyecto de Explotación de Información que se desarrolla en el contexto de una PyME.

- son respondidas por el ingeniero mediante el uso de valores lingüísticos (‘*nada’*, ‘*poco’*, ‘*regular’*, ‘*mucho’* y ‘*todo’*).

b. La formulación de un proceso que especifica las acciones a llevarse a cabo para determinar la viabilidad global del proyecto; dicho proceso está compuesto por cinco pasos, a saber:

- la transformación de los valores lingüísticos indicados por el ingeniero para cada condición (o pregunta) en intervalos difusos.
- el cálculo de la valoración de cada dimensión (*Plausibilidad*, *Adecuación* y *Éxito*) mediante una fórmula aplicada a los intervalos difusos correspondientes.
- el cálculo combinado de los valores de cada dimensión para determinar el valor asociado a la viabilidad global del proyecto
- una guía para interpretar todos los valores obtenidos y, de esta manera, identificar los puntos débiles asociados al desarrollo del proyecto.

II. Se ha determinado que es posible proponer un Modelo para la Estimación de Esfuerzo de proyectos de Explotación de Información dentro del ámbito de las PyMEs, el cual contiene:

a. La definición de ocho factores de costo asociados al proyecto que deben ser identificadas por el Ingeniero de Explotación de Información; dichos factores de costo presentan las siguientes particularidades:

- permiten determinar las características de la organización, de los datos disponibles y del proyecto.
- cabe destacar que estas características señaladas pueden conocerse al comienzo de un proyecto de Explotación de Información que se desarrolla en el contexto de una PyME.
- se encuentran agrupadas en tres tipos de factores (*Proyecto*, *Datos Disponibles* y *Recursos Disponibles*) de acuerdo al aspecto del proyecto se desea caracterizar.
- son respondidas por el ingeniero utilizando un valor numérico según la tabla correspondiente a cada factor de costo.

b. La formulación de un proceso que especifica las acciones que deben llevarse a cabo para llevar a cabo la estimación del esfuerzo requerido para desarrollarlo en forma

completa; dicho proceso cuenta con dos métodos de estimación que pueden aplicarse simultáneamente o en forma independiente:

- una fórmula lineal obtenida mediante regresión, la cual aplica en forma directa los valores de los factores de costo definidos por el ingeniero para calcular el esfuerzo (en meses/hombre).
- un método empírico, el cual combina los valores de los factores de costo definidos por el ingeniero para determinar el valor de cuatro coeficientes; que son posteriormente aplicados en una fórmula para calcular el esfuerzo (en meses/hombre).

Cabe recordar que la especificación de los modelos propuestos en el presente trabajo de tesis se ha focalizado en las características de los proyectos Explotación de Información desarrollados mediante la metodología CRISP-DM [Chapman *et al.*, 2000]. Asimismo, dado que los proyectos reales empleados en la validación fueron realizados aplicando dicha metodología, es posible considerar que ambos modelos son confiables para proyectos que se desarrollen en base a la misma. No obstante, y debido a que el Modelo de Proceso para Proyectos de Explotación de Información [Vanrell *et al.*, 2010; 2012] se basa en la metodología CRISP-DM, ambos modelos también pueden ser aplicados en proyectos que apliquen dicho Modelo de Proceso.

Teniendo en cuenta lo mencionado anteriormente, se considera de interés destacar las siguientes aportaciones anexas:

- Para cada modelo propuesto se ha identificado la tarea en la cual dicho modelo es susceptible de ser aplicado, tanto dentro de la metodología CRISP-DM, como del Modelo de Procesos para Explotación de Información.
- Para cada modelo propuesto se han identificado las condiciones que debe cumplir un proyecto de Explotación de Información bajo las cuales los resultados obtenidos pueden ser considerados válidos, en el marco las metodologías mencionadas anteriormente.

  Esas condiciones mencionadas en el punto anterior han sido identificadas mediante el análisis del comportamiento de cada modelo a partir de un enfoque estadístico; en tal sentido, dicho análisis se realiza en base a los datos de proyectos generados en forma pseudo-aleatoria a través del método de simulación Monte Carlo [Metropolis & Ulam, 1949; Kalos & Whitlock, 1986].
- Ambos modelos han sido implementados mediante planillas de cálculos para facilitar su utilización, las cuales se encuentran disponibles en [Pytel, 2013b].

6.2. FUTURAS LÍNEAS DE INVESTIGACIÓN

A lo largo del desarrollo de este trabajo de tesis han surgido cuestiones que, si bien no son centrales a los temas abordados en la misma, constituyen temas concomitantes que (en opinión del tesista) dan lugar a las siguientes futuras líneas de investigación:

I. En el caso del Modelo para la Evaluación de la Viabilidad de proyectos de Explotación de Información dentro del ámbito de las PyMEs, cabe citar las siguientes:

- **Estudiar el comportamiento del modelo en proyectos desarrollados con otras metodologías:**

  No obstante los proyectos que fueron utilizados en la validación del modelo se consideran representativos, se estima recomendable recolectar una mayor cantidad de proyectos que se hayan desarrollado dentro del marco de las PyMEs y, preferentemente, aplicando otras metodologías de desarrollo. De esta manera, se debe evaluar el comportamiento del modelo bajo nuevas condiciones, tales como, diferentes tipos de datos y nuevas clases de problemas de negocio, entre otros. A partir de dicho estudio, y con el objetivo de generalizar el modelo propuesto, se podría detectar la necesidad de ajustar el modelo agregando nuevas condiciones (o preguntas), o proponiendo nuevas fórmulas que enriquezcan el modelo.

- **Analizar los Riesgos asociados al Proyecto:**

  A partir de la interpretación de los resultados obtenidos por el modelo (valores y representación gráfica de los intervalos para cada dimensión) es posible identificar los puntos débiles que se encuentran asociados a los riesgos del proyecto; tales como fuentes de datos incompletos o desactualizados, mala documentación de la organización, falta de apoyo de los miembros de la organización, entre otros. Dado que la identificación y análisis de los riesgos del proyecto suele depender de la experiencia y conocimientos del ingeniero; se estima conveniente proponer un nuevo método que permita identificar dichos riesgos utilizando la información generada por el modelo. Asimismo, por cada riesgo se identificaría su probabilidad de ocurrencia, grado de criticidad y los planes de contingencia necesarios para mitigarlos.

II. En el caso del Modelo para la Estimación de Esfuerzo de proyectos de Explotación de Información dentro del ámbito de las PyMEs, cabe citar las siguientes:

- **Estudiar el comportamiento del modelo en proyectos desarrollados con otras metodologías:**

  No obstante los proyectos que fueron utilizados en la validación del modelo se consideran representativos, se estima recomendable recolectar una mayor cantidad de proyectos que se hayan desarrollado dentro del marco de las PyMEs y, preferentemente, aplicando otras metodologías de desarrollo. De esta manera, se debe evaluar el comportamiento del modelo bajo nuevas condiciones, tales como, diferentes tipos de organización y nuevas clases de problemas de negocio, entre otros. A partir de dicho estudio, y con el objetivo de generalizar el modelo propuesto, se podría detectar la necesidad de ajustar el modelo agregando nuevos factores de costo, introduciendo nuevos valores para cada factor de costo, o proponiendo nuevos métodos que enriquezcan el modelo.

- **Determinar la distribución del Esfuerzo por cada fase de la metodología:**

  Como resultado de la aplicación de los métodos incluidos en el modelo se obtiene el esfuerzo requerido para realizar el proyecto en forma completa (en meses/hombre). En tal sentido, se considera de utilidad adicionar nuevas fórmulas que permitan determinar, a partir del esfuerzo total ya calculado, la distribución del esfuerzo requerido para cada fase de la metodología utilizada. Es decir, la cantidad de tiempo que insumiría desarrollar cada una de las fases, lo cual haría más fácil realizar las tareas de planificación. Asimismo, se estima de interés contemplar la posibilidad de incluir otra fórmula que permita estimar la cantidad de personas que deben participar en cada una de las fases.

- **Ajustar el Esfuerzo estimado por cambios en el contexto del Proyecto:**

  Teniendo en cuenta que una vez que el desarrollo del proyecto ha comenzado, la estimación del esfuerzo puede quedar desactualizada en función de los cambios que por lo general tienen lugar en el contexto del mismo. En tal sentido de utilidad contar con un conjunto de fórmulas que permita re-estimar el esfuerzo requerido a los efectos de desarrollar el resto de las fases todavía no realizadas. Cabe señalar, estas nuevas fórmulas adicionales podrían utilizar la nueva información recolectada sobre el proyecto; así como también, el esfuerzo real que fue empleado en las fases ya realizadas.

7. REFERENCIAS

Ackoff, R. L. 1989. *From data to wisdom*. Journal of Applied Systems Analysis; 15, pp. 3–9.

Agarwal, R. y Kumar, M. 2001. *Estimating software projects*. Software Engineering Notes, 26(4), pp. 60-67.

Albrecht, A. J. y Gaffney Jr, J. E. 1983. *Software function, source lines of code, and development effort prediction: a software science validation*. Software Engineering, IEEE Transactions on, (6), pp. 639-648.

Álvarez, M. y Durán, J. 2009. *Manual de la Micro, Pequeña y Mediana Empresa. Una contribución a la mejora de los sistemas de información y el desarrollo de las políticas públicas*. San Salvador: CEPAL - Naciones Unidas.

Berger, J. O. 1985. *Statistical decision theory and Bayesian analysis*. New York: Springer.

Bielak, J. 2000. *Improving Size Estimates Using Historical Data*. IEEE Software, 17 (6), pp. 27-35, Nov./Dec. 2000, doi:10.1109/52.895165.

Bloch, A. 2003. *Murphy's law: The 26th Anniversary Edition*. Perigee Books.

Boehm, B. 1984. *Software Engineering Economics*. IEEE Transactions on Software Engineering, 10(1), pp. 4-21.

Boehm, B. W., Clark, Horowitz, Brown, Reifer, Chulani, Madachy, R. y Steece, B. 2000. *Software Cost Estimation with COCOMO II*. (1st ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

Bohan, W. F. 2003. *El poder oculto de la productividad: cómo mejorar la productividad en un 30% sin tener que despedir a nadie*. Editorial Norma.

Bolea, U., Jakličb, J., Papac, G. y Žabkard, J. 2011. *Critical Success Factors of Data Mining in Organizations*. Ljubljana.

Britos, P. y García-Martínez, R. 2009. *Propuesta de Procesos de Explotación de Información*. Proceedings XV Congreso Argentino de Ciencias de la Computación. Workshop de Base de Datos y Minería de Datos, pp. 1041-1050. ISBN 978-897-24068-4-1.

Chapman, P., Clinton, J., Keber, R., Khabaza , T., Reinartz, T., Shearer, C. y Wirth, R. 2000. *CRISP-DM 1.0 Step by step BI guide*. Edited by SPSS. Disponible en: http://tinyurl.com/crispDM

Charette, R.N. 2005. *Why software fails*. IEEE Spectrum 42(9), pp. 42-49. Disponible en: http://spectrum.ieee.org/computing/software/why-software-fails/0

Chen, Z., Menzies, T., Port, D. y Boehm, B. 2005. *Finding the right data for software cost modeling*. Software, IEEE, 22(6), Nov-Dec. 2005, pp. 38-46.

Cobos, C., Zuñiga, J., Guarin, J., León, E. y Mendoza, M. 2010. *CMIN-herramienta case basada en CRISP-DM para el soporte de proyectos de minería de datos*. Ingeniería e Investigación, 30(3), pp. 45-56.

Davenport, T.H. 2009. *Make Better Decisions*. Harvard Business Review, (November), pp. 117-123.

Edelstein, H. A. y Edelstein, H. C.. 1997. *Building, Using, and Managing the Data Warehouse*. Data Warehousing Institute, Prentice-Hall PTR, EnglewoodCliffs (NJ).

Fairley, R. E. 1992. *Recent advances in software estimation techniques*. Proceedings of the 14th International Conference on Software Engineering, pp. 382-391. ACM.

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P. 1996. *From data mining to knowledge discovery in databases*. AI magazine, 17(3), p. 37.

Fayyad, U.M. 2000. *Tutorial Report. Summer school of DM*. Monash University, Australia.

Hearst, M. 2003. *What Is Text Mining?* SIMS, UC Berkeley. Disponible en: http://people.ischool.berkeley.edu/~hearst/text-mining.html

García Martínez, R., Servente, M. y Pasquini, D. 2003. *Sistemas Inteligentes*. Editorial Nueva Librería. ISBN 987-1104-05-7.

García-Martínez, R. y Britos, P. 2004. *Ingeniería de Sistemas Expertos*. Editorial Nueva Librería. ISBN 987-1104-15-4.

García-Martínez, R., Britos, P., Pollo-Cattaneo, F., Rodriguez, D. y Pytel, P. 2011a. *Information Mining Processes Based on Intelligent Systems*. Proceedings of II International

Congress on Computer Science and Informatics (INFONOR–CHILE 2011), pp. 87–94. ISBN 978–956–7701–03–2.

García-Martínez, R., Lelli, R., Merlino, H., Cornachia, L., Rodriguez, D., Pytel, P. y Arboleya, H. 2011b. *Ingeniería de Proyectos de Explotación de Información para PYMES*. Proceedings XIII Workshop de Investigadores en Ciencias de la Computación, pp. 253-257. ISBN 978-950-673-892-1.

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

Gondar, J.E. 2005. *Metodología del Data Mining*. Number 84-96272-21-4. Data Mining Institute S.L..

Gómez, A., Juristo, N., Montes, C. y Pazos, J. 1997. *Ingeniería del Conocimiento*. Ed. Centro de Estudios Ramón Areces (Madrid).

Han, J. y Kamber, M. 2001. *Data mining: Concept and techniques*. Morgan Kaufmann Publishers, Inc.

Iglesias Pérez, M. C. (S/A) *Palabrario: Nivel de significación*.. Disponible en: http://www.educabarrie.org/palabrario/nivel-de-significacion

International Organization for Standardization (ISO). 2011. *ISO/IEC DTR 29110-1 Software Engineering - Lifecycle Profiles for Very Small Entities (VSEs) - Part 1: Overview*. International Organization for Standardization (ISO), Geneva, Switzerland.

Jang, J.S.R. 1997. *Fuzzy inference systems*. Upper Saddle River, NJ: Prentice-Hall.

JMACOE. 2013. *Cinco de los mejores software de minería de datos de Código Libre y Abierto*. Disponible en: http://blog.jmacoe.com/gestion_ti/base_de_datos/5-mejores-software-mineria-datos-codigo-libre-abierto/

Kalos, M.H. y Whitlock, P.A. 1986. *Monte Carlo Methods. Vol I. Basics*. John Wiley & Sons. New York. ISBN: 0471898392.

Kanungo, S. 2005. *Using Process Theory to Analyze Direct and Indirect Value-Drivers of Information Systems*. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, pp. 231-240.

KDnuggets. 2007. *Encuesta sobre metodologías utilizadas en Data Mining*. Disponible en: http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

Lakshmi, B. N., Raghunandhan, G. H. 2011. *Conceptual Overview of Data Mining*. Proceedings of the National Conference on Innovations in Emerging Technology 2011, pp. 27-32.

Laporte, C., Alexandre, S. & Renault, A. 2008. *Developing International Standards for VSEs*. IEEE Computer, 41(3): 98.

Lavravc, N., Motoda, H., Fawcett, T., Holte, R., Langley, P. y Adriaans, P. 2004. *Introduction: Lessons learned from data mining applications and collaborative problem solving*, Machine learning, vol. 57, nº 1, pp. 13-34.

López, D., García-Martínez R. y Marsiglio, A. 1991. *Factibilidad de Construcción de Sistemas Basados en Conocimiento*. Anales del II Simposio de Inteligencia Artificial y Robótica, Universidad Nacional de Luján (Buenos Aires), pp. 69-73.

Mansilla, D., Pollo, F., Britos, P., García-Martínez, R. 2013. *A Proposal of a Process Model for Requirements Elicitation in Information Mining Projects*. Lecture Notes in Business Information Processing, 139: 165-173. ISBN 978-3-642-36610-9.

Marbán, O., Menasalvas, E. y Fernández-Baizán, C. 2008. *A cost model to estimate the effort of data mining projects (DMCoMo)*. Information Systems, 33(1), pp. 133–150.

Marbán, O., Mariscal, G., y Segovia, J. 2009. *A Data Mining & Knowledge Discovery Process Model. Data Mining and Knowledge Discovery in Real Life Applications*. IN-TECH, 2009, p. 8.

Mariscal, G., Marbán, Ó., González, Á. L. y Segovia, J. 2007. *Hacia la Ingeniería de Data Mining: Un modelo de proceso para el desarrollo de proyectos*. II Congreso Español de Informática, pp. 139-148.

Mariscal, G., Marbán, O. y Fernández, C. 2010. A *survey of data mining and knowledge discovery process models and methodologies*. Knowledge Engineering Review, 25(2), pp. 137-166. doi:10.1017/S0269888910000032.

May, L.J. 1998. *Major causes of software project failures*. CrossTalk: The Journal of Defense Software Engineering. 11(6), pp. 9-12.

Metropolis, N. y Ulam, S. 1949. *The Monte Carlo Method*. Journal of the American Statistical Association; 44(247), pp. 335-341.

Nadali, A., Kakhky, E.N. y Nosratabadi, H.E. 2011. *Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system*. Electronics Computer Technology (ICECT). 3rd International Conference on Kanyakumari, Vol. 6, pp. 161–165. IEEE. doi: 10.1109/ICECTECH.2011.5942073.

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

Nemati, H. R. y Barko, C. D. 2003. *Key factors for achieving organizational data-mining success*. Industrial Management & Data Systems, 103(4), pp. 282-292. doi: 10.1108/02635570310470692.

Nie, G., Zhang, L., Liu, Y., Zheng, X. y Shi, Y. 2009. *Decision analysis of data mining project based on Bayesian risk*. Expert Systems with Applications, 36(3), pp. 4589–4594.

Organización para la Cooperación y el Desarrollo Económico (OECD). 2005. *Organization for Economic Cooperation and Development SME and Entrepreneurship Outlook 2005*. OECD Publishing.

Oktaba, H., Esquivel, C. A., Ramos, A. S., Martínez, A. M., Osorio, G. Q., López, M. R. y Lemus, M. Á. F. 2003. *Modelo de Procesos para la Industria de Software MoProSoft*. Secretaría de Economía de México.

Oktaba, H., Garcia, F., Piattini, M., Ruiz, F., Pino y F.J. y Alquicira, C. 2007. *Software Process Improvement: The COMPETISOFT Project*. Computer 40(10), pp. 21-28.

Guardía-Olmos, J., Freixa-Blanxart, M., Peró-Cebollero, M. y Turbany Oset, J. (2007). *Análisis de Datos en Psicología*. Delta Publicaciones.

Piatetsky-Shaphiro, G. y Frawley, W. 1991. *Knowledge Discovery in Databases*. AAAI/MIT Press, MA.

Pipino, L.L., Lee, Y.W. y Wang, R.Y. 2002. *Data quality assessment*. Communications of the ACM, 45(4), pp. 211–218.

Pohl, K. 1997. *Requirements Engineering: An Overview*. En M. Dekker, Encyclopedia of Computer Science and Technology, 36.

Pollo-Cattaneo, F., Britos, P., Pesado, P. y García-Martínez, R. 2010. *Proceso de Educción de Requisitos en Proyectos de Explotación de Información*. En Ingeniería de Software e Ingeniería del Conocimiento: Tendencias de Investigación e Innovación Tecnológica en Iberoamérica, pp. 1-11.

Pollo-Cattaneo, M., García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Rodríguez, D., Merlino, H., Pytel, P. y Vanrell, J. 2012. *Elementos para una Ingeniería de Explotación de Información*. Proyecciones 10(1), pp. 67–84. ISSN 1667–8400

Pollo-Cattaneo, M. F, Mansilla, D., Vegega, C, Pesado, P., García-Martínez, R. y Britos, P. 2013a. *Modelo de Procesos para la Gestión de Requerimientos en Proyectos de Explotación de Información*. Proceedings XIX Congreso Argentino de Ciencias de la Computación. ID 5618. ISBN 978-987-23963-1-2.

Pollo-Cattaneo, M.F., Mansilla, D., Vegega, C., Pytel, P., Pesado, P., García-Martínez, R. y Britos, P. 2013b. *Propuesta Integral de Manejo de Requerimientos en Proyectos de Explotación de Información*. En "Reflexiones sobre Ingeniería de Requisitos y Pruebas de Software" (Ed. Jaime Echeverri). Pág. 26-44. Editorial de la Corporación Universitaria Remington y Organización LACREST. ISBN 978-958-58070-3-7.

Pressman, R. 2005. *Ingeniería de Software: Un enfoque práctico*. 6ta edición, McGraw-Hill.

Putnam, L. H. 1978. *A general empirical solution to the macro software sizing and estimating problem*. Software Engineering, IEEE Transactions on, (4), pp.345-361.

Pyle, D. 1999. *Data preparation for data Mining*. Morgan Kaufmann.

Pyle, D. 2003. *Business Modeling and Business Intelligence*. Morgan Kaufmann.

Pytel, P., Tomasello, M., Rodríguez, D., Pollo-Cattaneo, F., Britos, P. y García-Martínez, R. 2011. *Estudio del Modelo Paramétrico DMCoMo de Estimación de Proyectos de Explotación de Información*. Proceedings XVII Congreso Argentino de Ciencias de la Computación, pp. 979–988. ISBN 978-950-34-0756-1.

Pytel, P. 2012a. *Implementación del Modelo de Viabilidad*. Disponible en: http://tinyurl.com/ViabPruConcepto

Pytel, P. 2012b. *Banco de Datos de Prueba generados por el Método Monte Carlo para analizar el Modelo de Viabilidad Propuesto*. Disponible en: http://tinyurl.com/BancoPruebaViab

Pytel, P. 2013a. *Banco de Datos de Prueba generados por el Método Monte Carlo para analizar el Modelo de Estimación Propuesto*. Disponible en: http://tinyurl.com/BancoPruebaEstim

Pytel, P. 2013b. *Implementación de los Modelos para Gestión de Proyectos*. Disponible en: http://tinyurl.com/implModelosTesisDoc

Ríos, M. D. 2006. *El Pequeño Empresario en ALC, las TIC y el Comercio Electrónico*. Instituto para la Conectividad en las Américas.

Rodríguez, D., Pollo-Cattaneo, F., Britos, P. y 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, pp. 664-673. ISBN 978-950-9474-49-9.

SAS Enterprise Miner. 2008. *SEMMA*. Disponible en: http://tinyurl.com/semmaSAS

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

Sim, J. 2003. *Critical success factors in data mining projects*. Ph.D. Thesis, University of North Texas.

Sommerville, I. 2005. *Ingeniería del software*. Pearson Educación.

Sommerville, I. y Sawyer, P. 1997. *Requirements Engineering: A Good Practice Guide*. Chichester, England: John Wiley & Sons.

Standish Group. 1995. *Chaos Report*. Disponible en: https://cs.nmt.edu/~cs328/reading/Standish.pdf

Standish Group. 2010. *Summary Report 2010*. Disponible en: http://blog.standishgroup.com/, http://insyght.com.au/special/2010CHAOSSummary.pdf

Strand, M. 2000. *The Business Value of Data Warehouses - Opportunities, Pitfalls and Future Directions*. Ph.D. Thesis, Department of Computer Science, University of Skovde.

Thomsen, E. 2003. *BI's Promised Land*. Intelligent Enterprise, 6(4), pp. 21-25.

Triola, M. 2013. *Estadística.* Pearson Educación, 11 edición.

Tukey, J. 1977. *Exploratory Data Analysis.* Addison-Wesley Publishing Company.

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

Vanrell, J., Bertone, R., García-Martínez, R. 2012. *Un Modelo de Procesos para Proyectos de Explotación de Información.* Proceedings Latin American Congress on Requirements Engineering and Software Testing, pp. 46-52. ISBN 978-958-46-0577-1.

Weisberg, S. 1985. *Applied Linear Regression*. Wiley & Sons, New York.

Wilcoxon, F. 1945. *Individual Comparisons by Ranking Methods*, Biometrics 1, pp. 80-83.

Yang, Q., Wu, X., Domingos, P., Elkan, C., Gehrke, J., Han, J., Heckerman, D., Keim, D., Liu, J., Madigan, D., Piatetsky-Shapiro, G., Raghavan, V., Rastogi, R., Stolfo, S., Tuzhilin, S., y Wah, B. 2006. *10 challenging problems in data mining research*. International Journal of Information Technology & Decision Making, 5(4): 597. doi: 10.1142/S0219622006002258

ANEXO A: PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN UTILIZADOS PARA DEFINIR EL MODELO DE ESTIMACIÓN DE ESFUERZO

En este anexo se indican los datos de los treinta y cuatro proyectos de explotación de información suministrados por colegas investigadores que fueron utilizados en la regresión aplicada para obtener la fórmula lineal (PEM_L) del Modelo de Estimación de Esfuerzo orientado para PyMEs.

Primero en las Tablas A.1, A.2 y A3 se indican el dominio de cada proyecto, el objetivo del negocio del proyecto y el esfuerzo real que fue necesario para desarrollarlo en forma completa. No es posible indicar mayor cantidad de información sobre cada proyecto por motivos de confidencialidad.

Posteriormente, en las Tablas A.4 a A.10 se indican la caracterización de los proyectos usando los factores de costo del modelo propuesto.

#	Dominio del Proyecto	Objetivos de Negocio	Esfuerzo Real (meses / hombre)
R1	Negocios / Industria	Se busca determinar cuáles son las necesidades que se buscan satisfacer en un auto la clase media.	6,08
R2	Servicio al Cliente	Como principales objetivos del negocio, se desea dar respuesta a dos aspectos esenciales del mismo: a. Aumentar la efectividad en el uso del presupuesto mensual destinado a promociones b. Disminuir la paulatina migración de clientes activos.	12,92
R3	Logística	Una empresa se dedica a la logística y distribución de productos de todo tipo para empresas privadas, tiene cuatro sucursales distribuidas entre Capital Federal y Gran Buenos Aires. El objetivo principal de la empresa es posicionarse como líder en transporte y logística privados para mercaderías privadas de otras empresas. Básicamente el mercado en el cual está inmersa se encuentra en constante crecimiento, debido a las ventas en Internet, lo cual conlleva a la necesidad de transportar mercaderías que se han comprado a distancia.	22,88
R4	Marketing	Se busca conocer cuáles son las características que determinan que una oportunidad de negocio sea lograda y las características que más influencia tienen en la perdida de una oportunidad de negocio.	5,54
R5	Marketing	Un banco desea armar campañas publicitarias dirigidas a sus distintos grupos de clientes con el objetivo de incrementar las ventas de sus productos. Para eso, necesitan determinar cuáles son los grupos de clientes y sus características.	10,94
R6	Servicio al Cliente	El problema de negocio planteado está íntimamente relacionado con la calidad de los servicios brindados. Cada cliente firma un acuerdo de servicio, comúnmente conocido como Service Level Agreement (SLA), de esta forma es responsabilidad de cumplir con las normas contractuales y proveer a sus clientes los servicios acordados. En el caso de no cumplir con estos acuerdos, la empresa será responsable por las pérdidas respondiendo económicamente ante la falla. Para evitar estas pérdidas se desean conocer las razones de los errores.	4,54
R7	Logística	Una empresa quiere determinar cuál es la forma óptima de almacenar los productos de los clientes dentro de las cámaras de frío, de modo que su búsqueda y salida sean más organizadas y predecibles. Y de esta manera reducir los costos que implica el tiempo de ubicación de los productos dentro de las cámaras, y el de salida. Así como también la posterior distribución de los productos, de manera rápida y eficiente.	11,50
R8	Servicio al Cliente	Es necesario identificar grupos de interés y proponer libros a cada grupo. Luego se hará un tratamiento especial por usuario filtrando aquellos títulos que ya haya clasificado (es decir que ya haya comprado / leído).	6,04
R9	Tecnología de la Información	El objetivo del negocio es aislar el correo basura, y de esta forma poder optimizar el uso de recursos y evitar la propagación del mismo.	5,96

Tabla A.1. Objetivo del Negocio y Esfuerzo real de los proyectos (R1 a R9).

#	Dominio del Proyecto	Objetivos de Negocio	Esfuerzo Real (meses / hombre)
R10	Marketing	El sector de marketing de una empresa desea incursionar en un ámbito de mercado determinado.	3,69
R11	Servicio al Cliente	Se desea mejorar la calidad de la atención brindada al cliente mediante la disminución del tiempo de respuesta, es decir, el tiempo que transcurre desde el momento en que se recibe el llamado del cliente solicitando el envío de un técnico hasta el momento en el que el técnico finaliza el diagnóstico sobre el problema del equipo. Como objetivo cuantificable se determinó que en un plazo máximo de 6 meses, un 60% de las órdenes de servicio deben ser resueltas en menos de 8 horas (equivalente a un día laboral).	8,94
R12	Recursos Humanos	Una empresa desea identificar las características coadyuvan al alto volumen de rotación de los empleados.	3,10
R13	Educación / Universidad	La Secretaría Académica de una Universidad desea diseñar nuevos cursos extracurriculares de especialización (arancelados), de acuerdo al interés de su población universitaria.	5,88
R14	Marketing	El objetivo de negocio es poder determinar si una persona pertenece a cierta clase social, para evaluar la posibilidad de que se convierta en un cliente para así poder realizar publicidad dirigida y prevenir gastos innecesarios.	6,88
R15	Negocios / Industria	La empresa desea poder evaluar qué modelos de automóviles son los más aptos para comprar según la aceptación de sus empleados.	2,13
R16	Educación / Pre-escolar	El caso de estudio se sitúa en la capital de Eslovenia y temporalmente en la década de 1980. En esta época hubo un exceso de inscripción a guarderías infantiles. Por lo que se debió realizar una selección entre las solicitudes recibidas dando una justificación objetiva a las mismas.	5,00
R17	Control de Transito	Identificar las características de los choques ocurridos durante el período 2006-2008 y poden identificar patrones que nos permitan tomar acciones preventivas para evitar las condiciones en que se favorece la ocurrencia de accidentes.	6,31
R18	Salud / Ambulancias	Una importante ART del mercado, está analizando la manera de reducir los costos de traslado de asegurados en tratamiento de un nosocomio a otro. Estos traslados se realizan desde la casa del asegurado hasta el prestador y en determinadas circunstancias se realizan en vehículos especialmente equipados, lo que encarece aún más los traslados.	5,31
R19	Logística	Una empresa dedicada a la elaboración de Fertilizantes cuenta con diferentes plantas de producción. Para el funcionamiento diario de dichas plantas se necesitan materiales de consumo, e incluso debe realizarse mantenimiento constante para garantizar el buen funcionamiento de las mismas. Para esta tarea es necesario contar con diferentes tipos de materiales e insumos (papelería, servicios, repuestos, lubricantes, etc.) que serán utilizados como repuestos de maquinarias e insumos para su mantenimiento y funcionamiento periódico. El principal objetivo es la fabricación de fertilizantes obteniendo el mayor rédito posible manteniendo los estándares de calidad, es por esto que se debe garantizar el buen funcionamiento ininterrumpido de sus plantas mediante un buen plan de mantenimiento.	8,88
R20	Tecnología de la Información	Optimización de estimaciones de complejidad de desarrollos: • Reducir en un 90% las discrepancias sobre la determinación de la complejidad del desarrollo. • Reducir en 30% el tiempo necesario para determinar la complejidad del desarrollo. • Poder realizar un estimado de cuál será la complejidad del desarrollo, antes de deber realizar el análisis necesario.	13,29
R21	Marketing	Una cadena de gimnasios desea determinar perfiles patrón de sus socios con el fin de ofrecerles servicios y productos exclusivos por perfil, como por ejemplo ropa o accesorios para determinadas disciplinas, viajes o excusiones para la tercera edad, campeonatos de futbol entre sedes, entre otros. El gimnasio cuenta con la base de datos de los clientes, donde se detallan sexo, año de nacimiento, actividades que realizan, frecuencia de asistencia, problemas físicos.	2,10
R22	Salud / Métodos Anticonceptivos	Una empresa desea conocer cuál de dos tipos de métodos anticonceptivos son los preferidos por las mujeres encuestadas para orientar sus investigaciones hacia este tipo de métodos. También desea conocer las características demográficas y socio económicas de cada uno de los grupos (los que eligen métodos anticonceptivos de corto plazo, de largo plazo y de los que no usan ningún método anti-conceptivo), en especial de aquellas que hayan seleccionado el método preferido, para futuras investigaciones, así como también para futuras campañas de marketing cuando se intente vender el producto. También se desea conocer la influencia que tiene la edad en la decisión del método anticonceptivo elegido.	14,54
R23	Marketing	Un banco ya tiene los usuarios categorizados en grupos y desea saber cuándo llega un nuevo cliente, que grupo asignarle.	2,60
R24	Servicio al Cliente	Un call center de soporte técnico a cliente ha detectado un aumento significativo en incidentes de un producto determinado y la empresa necesita saber cuál es la razón para este incremento. Se cuenta con la BD de incidentes con los datos del cliente y datos del incidente, entre estos, tipificación de la falla y tipificación de la solución.	11,87
R25	Control de Calidad	El área de control de calidad de una planta industrial desea distinguir los factores determinantes asociados a las fallas de las piezas que produce. La empresa posee datos sobre cada falla, y el estado de las piezas.	6,50
R26	Marketing	Establecer la mejor forma de armar packs de servicios secundarios en forma de promoción para conseguir la mayor cantidad de clientes posibles, de acuerdo a sus necesidades, dependiendo estos del plan primario al que pertenezcan y el sector en que se encuentren.	5,21

Tabla A.2. Objetivo del Negocio y Esfuerzo real de los proyectos (R10 a R26).

#	Dominio del Proyecto	Objetivos de Negocio	Esfuerzo Real (meses / hombre)
R27	Servicio al Cliente	Una compañía de venta de celulares dispone de una base de datos con datos de clientes de celular, con datos del tipo: si posee recarga con tarjeta o abono, las recargas que realiza un cliente, fecha de renovación del plan, saldo actual en su cuenta, entre otras. Con base en esta información quiere saber cuándo viene un cliente nuevo, que celular ofrecerle primero, y que ventajas le puede ofrecer con ese celular.	2,43
R28	Control de Transito	El objetivo del Gobierno Provincial es proveer asistencia vial para ordenar el tránsito en las situaciones en que el clima puede provocar dificultades, como visibilidad reducida o calzada resbaladiza.	4,26
R29	Negocios / Industria	Mejorar las metodologías utilizadas por el área de producción de vinos a partir de las fórmulas obtenidas en la implementación del proyecto y obtener diferentes tipos de productos de acuerdo a su calidad (buena, muy buena, excelente).	7,21
R30	Control de Calidad	El área de control de calidad de una planta industrial desea distinguir los factores determinantes asociados a las fallas de las piezas que produce. La empresa posee datos sobre cada falla y el estado de las piezas.	8,76
R31	Servicio al Cliente	Preservar y fidelizar a los clientes, aumentar la cobertura de los servicios y mejorar la tecnología.	16,25
R32	Marketing	Una empresa mayorista busca alternativas para generar promociones que puedan generar más ventas.	10,67
R33	Control de Transito	El objetivo de negocio es reducir la severidad de los accidentes mediante el análisis de sus causas.	10,75
R34	Social	Se desea implementar acciones concretas, idóneas e imprescindibles que hagan factibles mejorar las posibilidades y condiciones para la reinserción social de los internos, a fin que adquieran a su egreso la capacidad de desempeñarse como seres socialmente útiles. Por otro lado, también se desea implementar nuevos programas de tratamiento individualizados, orientados a brindar oportunidades de cambio en las conductas de los internos.	10,17

Tabla A.3 Objetivo del Negocio y Esfuerzo real de los proyectos (R27 a R34).

#	Tipo de objetivo de explotación de información (OBTY)	Grado de apoyo de los miembros de la organización (LECO)	Cantidad y tipo de los repositorios de datos disponibles (AREP)	Cantidad de tuplas disponibles en la tabla principal (QTUM)	Cantidad de tuplas disponibles en tablas auxiliares (QTUA)	Nivel de conocimiento sobre los datos (KLDS)	Nivel de conocimiento y experiencia del equipo de trabajo (KEXT)	Funcionalidad de las herramientas disponibles (TOOL)
R1	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 20.001 y 80.000 tuplas en la tabla principal (4).	No se utilizan tablas auxiliares (1).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R2	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Hasta 100 tuplas en la tabla principa l (1).	No se utilizan tablas auxiliares (1).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R3	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (4).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 101 y 1.000 tuplas en la tabla principal (2).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Las tablas y repositorios no están documentados y existen expertos en los datos pero no están disponibles para explicarlos (5).	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos (1).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R4	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Hasta 100 tuplas en la tabla principal (1).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R5	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 20.001 y 80.000 tuplas en la tabla principal (4).	No se utilizan tablas auxiliares (1).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).

Tabla A.4. Caracterización de los proyectos usando los factores de costo (R1 a R5).

#	Tipo de objetivo de explotación de información (OBTY)	Grado de apoyo de los miembros de la organización (LECO)	Cantidad y tipo de los repositorios de datos disponibles (AREP)	Cantidad de tuplas disponibles en la tabla principal (QTUM)	Cantidad de tuplas disponibles en tablas auxiliares (QTUA)	Nivel de conocimiento sobre los datos (KLDS)	Nivel de conocimiento y experiencia del equipo de trabajo (KEXT)	Funcionalidad de las herramientas disponibles (TOOL)
R6	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 101 y 1.000 tuplas en la tabla principal (2).	Más de 50.000 tuplas en las tablas auxiliares (4).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos (3).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R7	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Sólo 1 repositorio disponible (1).	Hasta 100 tuplas en la tabla principal (1).	No se utilizan tablas auxiliares (1).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R8	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos (1).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R9	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R10	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R11	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 20.001 y 80.000 tuplas en la tabla principal (4).	No se utilizan tablas auxiliares (1).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).

Tabla A.5. Caracterización de los proyectos usando los factores de costo (R6 a R11).

#	Tipo de objetivo de explotación de información (OBTY)	Grado de apoyo de los miembros de la organización (LECO)	Cantidad y tipo de los repositorios de datos disponibles (AREP)	Cantidad de tuplas disponibles en la tabla principal (QTUM)	Cantidad de tuplas disponibles en tablas auxiliares (QTUA)	Nivel de conocimiento sobre los datos (KLDS)	Nivel de conocimiento y experiencia del equipo de trabajo (KEXT)	Funcionalidad de las herramientas disponibles (TOOL)
R12	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (4).	Entre 2 y 4 repositorios con tecnología no compatible para la integración (3).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R13	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Hasta 100 tuplas en la tabla principal (1).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos (3).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R14	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos (3).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R15	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Hasta 100 tuplas en la tabla principal (1).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R16	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 20.001 y 80.000 tuplas en la tabla principal (4).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).

Tabla A.6. Caracterización de los proyectos usando los factores de costo (R12 a R16).

#	Tipo de objetivo de explotación de información (OBTY)	Grado de apoyo de los miembros de la organización (LECO)	Cantidad y tipo de los repositorios de datos disponibles (AREP)	Cantidad de tuplas disponibles en la tabla principal (QTUM)	Cantidad de tuplas disponibles en tablas auxiliares (QTUA)	Nivel de conocimiento sobre los datos (KLDS)	Nivel de conocimiento y experiencia del equipo de trabajo (KEXT)	Funcionalidad de las herramientas disponibles (TOOL)
R17	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 20.001 y 80.000 tuplas en la tabla principal (4).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos (3).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R18	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Más de 5 repositorios con tecnología compatible para la integración (4).	Entre 101 y 1.000 tuplas en la tabla principal (2).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos (3).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R19	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (5).	Más de 5 repositorios con tecnología compatible para la integración (4).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R20	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre la identificación de una clase desconocida previamente (5).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (4).	Más de 5 repositorios con tecnología no compatible para la integración (5).	Entre 20.001 y 80.000 tuplas en la tabla principal (4).	No se utilizan tablas auxiliares (1).	Las tablas y repositorios no están documentados y existen expertos en los datos pero no están disponibles para explicarlos (5).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R21	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).

Tabla A.7. Caracterización de los proyectos usando los factores de costo (R17 a R21).

#	Tipo de objetivo de explotación de información (OBTY)	Grado de apoyo de los miembros de la organización (LECO)	Cantidad y tipo de los repositorios de datos disponibles (AREP)	Cantidad de tuplas disponibles en la tabla principal (QTUM)	Cantidad de tuplas disponibles en tablas auxiliares (QTUA)	Nivel de conocimiento sobre los datos (KLDS)	Nivel de conocimiento y experiencia del equipo de trabajo (KEXT)	Funcionalidad de las herramientas disponibles (TOOL)
R22	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos (1).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R23	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Entre 2 y 4 repositorios con tecnología no compatible para la integración (3).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R24	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 101 y 1.000 tuplas en la tabla principal (2).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R25	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R26	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Entre 2 y 4 repositorios con tecnología no compatible para la integración (3).	Entre 101 y 1.000 tuplas en la tabla principal (2).	No se utilizan tablas auxiliares (1).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).

Tabla A.8. Caracterización de los proyectos usando los factores de costo (R22 a R26).

#	Tipo de objetivo de explotación de información (OBTY)	Grado de apoyo de los miembros de la organización (LECO)	Cantidad y tipo de los repositorios de datos disponibles (AREP)	Cantidad de tuplas disponibles en la tabla principal (QTUM)	Cantidad de tuplas disponibles en tablas auxiliares (QTUA)	Nivel de conocimiento sobre los datos (KLDS)	Nivel de conocimiento y experiencia del equipo de trabajo (KEXT)	Funcionalidad de las herramientas disponibles (TOOL)
R27	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre la identificación de una clase desconocida previamente (5).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Sólo 1 repositorio disponible (1).	Hasta 100 tuplas en la tabla principal (1).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos (1).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R28	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R29	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Más de 5 repositorios con tecnología compatible para la integración (4).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R30	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	No se utilizan tablas auxiliares (1).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos (1).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
R31	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre la identificación de una clase desconocida previamente (5).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (4).	Sólo 1 repositorio disponible (1).	Entre 20.001 y 80.000 tuplas en la tabla principal (4).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).

Tabla A.9. Caracterización de los proyectos usando los factores de costo (R27 a R31).

#	Tipo de objetivo de explotación de información (OBTY)	Grado de apoyo de los miembros de la organización (LECO)	Cantidad y tipo de los repositorios de datos disponibles (AREP)	Cantidad de tuplas disponibles en la tabla principal (QTUM)	Cantidad de tuplas disponibles en tablas auxiliares (QTUA)	Nivel de conocimiento sobre los datos (KLDS)	Nivel de conocimiento y experiencia del equipo de trabajo (KEXT)	Funcionalidad de las herramientas disponibles (TOOL)
R32	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (4).	Entre 2 y 4 repositorios con tecnología no compatible para la integración (3).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R33	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (4).	Sólo 1 repositorio disponible (1).	Hasta 100 tuplas en la tabla principal (1).	No se utilizan tablas auxiliares (1).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos (3).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
R34	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	No se utilizan tablas auxiliares (1).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).

Tabla A.10. Caracterización de los proyectos usando los factores de costo (R32 a R34).

ANEXO B: PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN UTILIZADOS PARA VALIDAR LOS MODELOS

En este anexo se indican los datos de los treinta y siete proyectos de explotación de información suministrados por colegas investigadores. Estos proyectos son utilizados durante la validación de los Modelos de Evaluación de la Viabilidad y Estimación de Esfuerzo propuestos en el capítulo 5 de este trabajo de tesis.

Primero, en las Tablas B.1, B.2, B.3 y B.4 se indican el dominio y el objetivo del negocio para cada uno de dichos proyecto junto con el proceso de explotación de información que fue aplicado. No es posible indicar mayor cantidad de información sobre cada proyecto por motivos de confidencialidad.

Posteriormente, en la Tabla B.5 se indica para cada proyecto información sobre el estado final del proyecto, su esfuerzo real, la valoración de las dimensiones y su viabilidad global.

Finalmente, los proyectos son caracterizados según el modelo de la viabilidad propuesto (Tabla B.6), y los factores de costo del modelo de estimación propuesto (Tablas B.7 a B.12).

#	Dominio del Proyecto	Objetivos de Negocio	Proceso de Explotación de Información
P1	Negocios / Industria	El objetivo del negocio es clasificar los diferentes tipos de productos y ver la aceptación de las personas en cuanto a los mismos, así como también qué características tiene el producto con mayor aceptación.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P2	Marketing	Debido a que en nuestro país existe un gran incremento en el segmento medio, la compañía tiene como objetivo ganar mercado captando nuevos clientes de gama media. Para eso, necesitan determinar cuáles son las necesidades que buscan satisfacer ese nicho de mercado.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P3	Negocios / Industria	Los altos directivos de una empresa desean aumentar la presencia y ampliar el mercado. Lo ideal es el lanzamiento de un producto nuevo que sea un éxito en ventas en un mediano plazo. El nuevo concepto será proclamado como una nueva unidad destinada a producción la cual creará más empleo, más ventas y por ende más ingresos.	Se utiliza el proceso de descubrimiento de reglas de comportamiento y ponderación de interdependencia de atributos.
P4	Marketing	Se busca la manera de identificar el comportamiento de los clientes, para poder entender qué tipo de cliente está más predispuesto a la adquisición de algún paquete de productos. El objetivo buscado es incrementar el nivel de aceptación y de ventas de paquetes de producto.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P5	Servicio al Cliente	Los objetivos del proyecto son dar publicidad personalizada a los usuarios, ubicar los anuncios en las secciones más óptimas.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P6	Salud / Análisis Enfermedades	Realizar un análisis de las causas que contribuyen a que los bebés presenten determinadas enfermedades al nacer, en un contexto de alto índice de indigencia, bajo nivel educativo, madres adolescentes e ingresos muy bajos.	Se utiliza descubrimiento de reglas de comportamiento y ponderación de los atributos detectados.

Tabla B.1. Datos generales de los proyectos (P1 a P6).

#	Dominio del Proyecto	Objetivos de Negocio	Proceso de Explotación de Información
P7	Servicio al Cliente	El sector de Mesa de Ayuda de una entidad gubernamental utiliza un sistema para registrar cada llamado que se recibe. De esa manera se puede identificar un pedido de reparación, cambio o mal funcionamiento de alguna computadora, o de alguno de sus periféricos que utiliza para poder luego asignar un técnico que procederá a resolver el inconveniente.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P8	Servicio al Cliente	La finalidad es mejorar la imagen que tienen los clientes de la empresa al brindarles un mejor servicio de distribución de los pedidos. Esto significa hallar los factores tanto internos como externos (es decir, relacionados con las distribuidoras tercerizadas) de la empresa que inciden en el retraso de los pedidos a entregar a los clientes.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P9	Negocios / Industria	La finalidad es reunir a los mejores talentos y capitales, esforzándose para lograr las mejores tecnologías mundiales avanzadas, la posesión de derechos propietarios intelectuales independientes, la creación de una marca internacionalmente famosa, el desarrollo del mercado automovilístico mundial y la clase mundial, entre los fabricantes automovilísticos superiores.	Se utilizan las técnicas de descubrimiento de grupos y descubrimiento de reglas de comportamiento. Luego se ponderan que características de las reglas representativas.
P10	Negocios / Industria	Se ha decidido identificar los atributos claves que hacen tener a los vinos una gran calidad. Se pretende que una vez detectados estos atributos, los mismos se mejoren en los vinos de menor calidad.	Se utiliza el proceso de descubrimiento de reglas de comportamiento y ponderación de interdependencia de atributos.
P11	Salud / Obra Social	Una Obra Social desea calificar a sus afiliados y adaptar planes para cada tipo de forma que ayude a la mejora en la rentabilidad de la empresa.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P12	Negocios / Industria	Una empresa de agro-insumos desea analizar el comportamiento de sus ventas y clientes dependiendo de la época del año. Para ello se desea determinar cuáles son los artículos más vendidos en las distintas épocas del año teniendo en cuenta su precio para poder mantener un stock adecuado y cuáles son las épocas de año en la cuales se registran las mayores ventas.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P13	Servicio al Cliente	Una empresa fundada en Estados Unidos, enfocada en el desarrollo de sitios web personalizables para los fanáticos de NASCAR, desea mejorar las características de sus sitios web y poder captar así mayor cantidad de clientes.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P14	Marketing	Una zapatería desea mejorar las ventas, promociones y ofertas realizadas a los clientes a partir de un mejor conocimiento de sus hábitos de consumo.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P15	Salud / Atención Clínica	Un centro de salud desea brindar alivio y contención a los afiliados de PAMI de la manera más eficaz y eficiente posible, para mejorar su calidad de vida	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P16	Servicio al Cliente	Una librería desea establecer un sistema de descuentos para las compras realizadas desde la web basado en categorías definidas a partir de los distintos tipos de clientes.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P17	Servicio al Cliente	Un banco busca conocer en detalle los patrones de interacción de los clientes con la entidad para poder adelantarse al comportamiento de los mismos y encontrar nuevas oportunidades de oferta de productos o aplicar estrategias de derivación. Se llama derivación a la traslación de las interacciones de un cliente desde un canal costoso, como lo es la caja por requerir un empleado que atienda, hacia un canal menos costoso, como Internet. Teniendo esto, el área de marketing podrá focalizar de una manera mucho más óptima sus estrategias, maximizando las ganancias y reduciendo los costos, evitando envíos de publicidad innecesarios a clientes que no les interesará determinado producto.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P18	Servicio al Cliente	Una empresa de seguro desea incrementar la cantidad de seguros personales cotizados por parte de sus clientes mediante la diferenciación de la competencia y ofrecer un valor agregado a sus clientes.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P19	Negocios / Industria	Una compañía de esterilización de productos farmacéuticos desea determinar la forma más óptima de realizar la esterilización dado la importancia en la calidad de la esterilización de materiales y en la reducción de costos como todo proceso industria.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.

Tabla B.2. Datos generales de los proyectos (P7 a P19).

#	Dominio del Proyecto	Objetivos de Negocio	Proceso de Explotación de Información
P20	Servicio al Cliente	Una importante empresa posee un sitio de alquileres de películas online. La misma posee una amplia recolección de datos referidos a calificaciones de una amplia variedad de usuarios. Para ello, es necesario identificar grupos de interés y proponer nuevos alquileres de películas a cada grupo. Luego se hará un tratamiento especial por usuario filtrando aquellos títulos que ya haya clasificado	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos y luego ponderación de interdependencia de atributos.
P21	Negocios / Industria	Una de las redes de inmobiliarias de Capital Federal, que tiene fuerte presencia en el barrio de Almagro, tiene el objetivo de comprender el contexto de su situación actual en cuanto a los tiempos de venta de inmuebles.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P22	Servicio al Cliente	Una empresa de tarjeta de créditos posee como objetivo la ubicación y consolidación de los distintos productos y servicios crediticios, ofrecidos tanto a través de las entidades bancarias existentes en el país, cómo por medio de su propia organización. El problema a abordar será la venta de un servicio con poca demanda por los clientes, el cual es denominado Card to Card (C2C). El mismo consiste en la transferencia de saldo de una tarjeta de crédito a otra tarjeta de crédito, lo cual genera un incremento del disponible de la cuenta a la cual se asocia el plástico receptor, y un decremento del disponible de la cuenta a la cual pertenece el plástico emisor.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P23	Marketing	En una droguería se plantea la posibilidad de comprar ciertos productos médicos en altas cantidades para obtener un mejor precio de compra, situación que permitirá ofrecer a los clientes productos en promociones y así incrementar los ingresos.	Se utiliza el proceso de descubrimiento de reglas de comportamiento y luego ponderación de interdependencia de atributos.
P24	Educación / Universidad	Desde la Secretaria Académica de una Universidad se desea mejorar el seguimiento de los alumnos de las carreras que se dictan en la facultad a partir de la caracterización de los alumnos desertores y grupos de riesgo, con el fin de realizar actividades de prevención para tratar los factores con influencia en el rendimiento académico y la deserción de los alumnos	Se utiliza el proceso de descubrimiento de reglas de comportamiento y luego ponderación de interdependencia de atributos.
P25	Educación / Primaria y Secundaria	Se desean estudiar los datos obtenidos del INDEC en los censos del 2001 y 2010 con la intención de encontrar relaciones entre indicadores socioeconómicos y la educación. Se quiere encontrar criterios y variables que afecten positivamente a la hora de destinar fondos, para mejorar la eficiencia y eficacia del proceso de inversión.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P26	Marketing	Una empresa de insumos agropecuarios desea ampliar su mercado para la venta de este tipo de insumos para lo cual desea obtener tendencias sobre los tipos de productos más vendidos en los diferentes trimestres del año y sus características más importantes.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P27	Salud / Obra Social	Con el propósito de ampliar su clientela, una Obra Social sindical busca cubrir la totalidad del monto de los reintegros, lo cual le ha provocado un desbalance financiero. A partir de esto, es que desea mejorar la toma de decisiones sobre la aprobación o rechazo de los reintegros.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P28	Salud / Obra Social	Una obra social ubicada en la localidad de Concepción del Uruguay (Entre Ríos) cuenta con 17.000 afiliados siendo una de las más numerosas de la ciudad. Quiere poder implementar estrategias de negocio en la empresa para la organización de los mismos.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P29	Educación / Bioblioteca	Una biblioteca aspira a ser el referente como biblioteca con prestigio y renombre en temas de ingeniería, gestión, medio ambiente, educación y cultura en la región al entender en la organización, planeamiento y funcionamiento técnico-administrativo.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P30	Negocios / Industria	Una empresa industrial argentina dedicada al diseño, fabricación y comercialización de acoplados, semirremolques, furgones térmicos y carrocerías para camiones, desea mejorar su línea de ensamble de unidades del área Producción haciéndola más eficaz y eficiente.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos y luego ponderación de interdependencia de atributos.
P31	Educación / Universidad	Los docentes de la asignatura de una asignatura universitaria de grado desean estudiar las distintas alternativas en el desarrollo de las clases para poder aumentar el porcentaje de aprobación de los exámenes parciales, haciendo hincapié en los temas que más complican a los alumnos.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.

Tabla B.3. Datos generales de los proyectos (P20 a P31).

#	Dominio del Proyecto	Objetivos de Negocio	Proceso de Explotación de Información
P32	Educación / Universidad	La cátedra de la asignatura de una asignatura universitaria de grado ha exteriorizado su interés por saber, tomando como base los exámenes finales correspondientes al último año, cuáles son las preguntas más relevantes a la hora de determinar si se aprueba o no el final. Al identificar estos temas, es posible hacer hincapié durante el dictado de la asignatura para poder mejorar el porcentaje de alumnos aprobados.	e utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P33	Turismo	Se desean estudiar los datos obtenidos del INDEC en el censo 2010 con la intención de encontrar patrones en el comportamiento de turistas en la zona del literal argentino. Se quiere encontrar criterios asociados a la ocupación hotelera y movilidad de los turistas en distintas épocas del año.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.
P34	Salud / Enfermedades	Se desea analizar la información recolectada en el repositorio de información médica NCBI sobre el genotipo de pacientes que presentan diferentes tipos de enfermedades.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P35	Educación / Universidad	Una universidad ha recopilado datos sobre los cursos dados por ayudantes de cátedra en tres semestres regulares y dos semestres de verano. Se recibieron evaluaciones de performance de estos ayudantes que se consideran para determinar su performance en los cursos dictados.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P36	Educación / Universidad	Detectar la correcta derivación de fondos por parte de las jurisdicciones responsables sobre los fondos transferidos por el Gobierno Nacional en referencia al Fondo Nacional de Incentivo Docente (FONID) durante el año 2006, que incluyeron beneficios devengados en el año 2005.	Se utiliza el proceso de descubrimiento de reglas de pertenencia a grupos.
P37	Turismo	Una empresa de turismo desea mejorar su oferta de paquetes turísticos teniendo en cuenta el poder adquisitivo, es decir la proyección de gastos de sus clientes. La expectativa de la empresa a largo plazo es realizar es promociones de turismos a clientes nacionales e internacionales de alta gama.	Se utiliza el proceso de descubrimiento de reglas de comportamiento.

Tabla B.4. Datos generales de los proyectos (P32 a P37).

#	Estado Final del Proyecto	Esfuerzo Real[1] (meses / hombre)	Valoración de la Plausibilidad[2]	Valoración de la Adecuación[2]	Valoración del Éxito[2]	Valoración de la Viabilidad[3]
P1	*Finalizado Satisfactoriamente*	2,41	8	7	4	6,33
P2	*Finalizado Satisfactoriamente*	7,00	7	6	5	6,00
P3	*Finalizado Satisfactoriamente*	1,64	8	5	6	6,33
P4	*Finalizado Satisfactoriamente*	3,65	6	6	4	5,33
P5	*Finalizado Satisfactoriamente*	9,35	6	8	7	7,00
P6	*Finalizado Satisfactoriamente*	11,63	6	5	5	5,33
P7	*Finalizado Satisfactoriamente*	6,73	5	5	5	5,00
P8	*Finalizado Satisfactoriamente*	5,40	6	5	6	5,67
P9	*Finalizado Satisfactoriamente*	8,38	7	6	6	6,33
P10	*Finalizado Satisfactoriamente*	1,56	6	5	6	5,67
P11	*Finalizado Satisfactoriamente*	9,70	8	5	6	6,33
P12	*Finalizado Satisfactoriamente*	5,24	7	8	7	7,33
P13	*Finalizado Satisfactoriamente*	5,00	7	5	6	6,00
P14	*Finalizado Satisfactoriamente*	8,97	7	7	6	6,67
P15	*Finalizado Satisfactoriamente*	2,81	9	7	8	8,00
P16	*Finalizado Satisfactoriamente*	11,80	7	6	5	6,00
P17	*Finalizado Satisfactoriamente*	2,79	6	5	5	5,33
P18	*Finalizado Satisfactoriamente*	3,88	5	5	6	5,33
P19	*Finalizado Satisfactoriamente*	5,70	8	7	7	7,33
P20	*Finalizado Satisfactoriamente*	8,54	9	7	5	7,00
P21	*Finalizado Satisfactoriamente*	10,61	8	6	5	6,33
P22	*Finalizado Satisfactoriamente*	6,88	7	6	6	6,33
P23	*Finalizado Satisfactoriamente*	11,20	7	7	8	7,33
P24	*Finalizado Satisfactoriamente*	9,70	7	8	5	6,67
P25	*Finalizado Satisfactoriamente*	7,30	5	7	5	5,67
P26	*Finalizado Satisfactoriamente*	5,31	8	8	8	8,00
P27	*Finalizado Satisfactoriamente*	6,10	8	6	7	7,00
P28	*Finalizado Satisfactoriamente*	10,00	8	6	7	7,00
P29	*Finalizado Satisfactoriamente*	6,43	7	5	7	6,33
P30	*Finalizado Satisfactoriamente*	9,80	8	8	6	7,33
P31	*Finalizado Satisfactoriamente*	1,50	7	6	8	7,00
P32	*Finalizado Satisfactoriamente*	3,78	7	7	8	7,33
P33	*Cancelado*	-	3	4	3	3,33
P34	*Cancelado*	-	4	5	2	3,67
P35	*Cancelado*	-	3	4	3	3,33
P36	*Cancelado*	-	5	3	2	3,33
P37	*Cancelado*	-	4	2	1	2,33

Tabla B.5. Información complementaria sobre los proyectos para ser utilizada en la validación de los modelos.

Notas de la tabla B.5:

([1]) Esfuerzo real requerido para realizar el proyecto en forma completa; por lo que este valor sólo se encuentra definido sólo para los proyectos finalizados satisfactoriamente (que no fueron cancelados antes de ser terminados).

([2]) Para definir la valoración de cada dimensión (plausibilidad, adecuación y éxito), se les pidió a investigadores expertos en el dominio que respondan una encuesta con las siguientes preguntas sobre el proyecto:

a) En su opinión ¿con qué grado de dificultad fue posible realizar el proyecto?
b) Para usted, ¿en qué medida la aplicación de los algoritmos de explotación de información permitió resolver el problema de negocio?
c) Teniendo en cuenta los resultados del proyecto ¿cuán exitosa fue la realización del proyecto?

Cada pregunta sólo podía ser respondida con un valor de 1 a 10, donde 10 se considera como el mayor valor y 1 el menor. En el caso de la Plausibilidad se asigna el valor contrario (1 en lugar de 10, 2 en lugar de 8, 3 en lugar de 7 y así sucesivamente) para representar la valoración de esta dimensión.

([3]) Para definir la valoración de la viabilidad del proyecto se han promediado los valores asignados para cada dimensión.

#	P1	P2	A1	A2	A3	E1	P3	A4	A5	E2	E3	P4	E4
P1	mucho	mucho	mucho	regular	mucho	poco	mucho	regular	mucho	mucho	regular	mucho	mucho
P2	todo	regular	todo	regular	regular	mucho	mucho	mucho	poco	mucho	regular	mucho	poco
P3	poco	todo	mucho	regular	regular	regular	mucho	regular	mucho	mucho	mucho	mucho	poco
P4	mucho	regular	regular	todo	mucho	todo	poco	regular	todo	nada	todo	mucho	mucho
P5	mucho	regular	mucho	todo	regular	mucho	poco	todo	todo	regular	mucho	mucho	todo
P6	regular	regular	mucho	regular	mucho	mucho	mucho	regular	regular	mucho	regular	regular	poco
P7	regular	regular	mucho	mucho	regular	regular	mucho	regular	regular	regular	regular	regular	mucho
P8	mucho	regular	todo	regular	mucho	mucho	mucho	todo	poco	poco	regular	mucho	todo
P9	mucho	mucho	mucho	regular	mucho	regular	mucho	regular	regular	mucho	regular	mucho	regular
P10	regular	regular	todo	regular	regular	regular	mucho	regular	regular	mucho	regular	mucho	regular
P11	mucho	regular	mucho	todo	todo	mucho	regular	regular	regular	regular	regular	todo	regular
P12	todo	mucho	todo	todo	todo	mucho	mucho	regular	mucho	regular	mucho	mucho	mucho
P13	mucho	regular	mucho	poco	regular	regular	regular	mucho	regular	mucho	todo	mucho	mucho
P14	regular	regular	mucho	regular	regular	mucho	todo	mucho	todo	poco	todo	mucho	mucho
P15	todo	todo	todo	regular	mucho	mucho	todo	mucho	mucho	mucho	todo	mucho	mucho
P16	todo	todo	mucho	regular	todo	mucho	poco	regular	mucho	poco	mucho	mucho	todo
P17	mucho	mucho	regular	todo	poco	mucho	regular	mucho	mucho	regular	regular	regular	regular
P18	regular	mucho	mucho	regular	mucho	regular	regular	mucho	poco	regular	regular	mucho	mucho
P19	regular	todo	todo	regular	mucho	todo	mucho	mucho	mucho	poco	todo	mucho	regular
P20	todo	todo	todo	todo	todo	mucho	mucho	regular	regular	regular	todo	mucho	poco
P21	todo	mucho	todo	poco	mucho	todo	mucho	todo	mucho	poco	mucho	todo	regular
P22	mucho	regular	poco	todo	poco	mucho	mucho	mucho	todo	mucho	regular	mucho	mucho
P23	regular	mucho	regular	regular	mucho	mucho	regular	regular	regular	mucho	mucho	regular	regular
P24	mucho	todo	regular	regular	todo	mucho	mucho	todo	regular	todo	todo	poco	mucho
P25	mucho	todo	poco	todo	todo	regular	todo	mucho	poco	todo	mucho	poco	mucho
P26	mucho	todo	todo	mucho	mucho	regular	regular	mucho	regular	poco	regular	mucho	mucho
P27	regular	todo	mucho	mucho	mucho	mucho	regular	mucho	mucho	regular	regular	mucho	regular
P28	regular	todo	regular	poco	todo	mucho	mucho	mucho	regular	mucho	regular	poco	regular
P29	todo	todo	regular	mucho	mucho	mucho	regular	mucho	regular	mucho	regular	mucho	todo
P30	regular	mucho	todo	mucho	mucho	mucho	mucho	poco	poco	mucho	mucho	regular	regular
P31	mucho	todo	regular	regular	mucho	mucho	mucho	mucho	mucho	regular	mucho	mucho	mucho
P32	mucho	todo	regular	regular	mucho	regular	mucho	mucho	regular	regular	mucho	todo	mucho
P33	poco	todo	poco	regular	todo	regular	regular	regular	poco	todo	mucho	poco	poco
P34	poco	todo	todo	poco	mucho	poco	regular	poco	nada	mucho	mucho	nada	poco
P35	regular	todo	regular	regular	regular	poco	regular	regular	nada	regular	mucho	todo	regular
P36	regular	mucho	regular	poco	poco	poco	poco	poco	mucho	poco	mucho	regular	regular
P37	regular	nada	regular	poco	nada	mucho	regular	poco	regular	regular	regular	regular	regular

Tabla B.6. Caracterización de los proyectos para el modelo de evaluación de la viabilidad.

#	OBTY	LECO	AREP	QTUM	QTUA	KLDS	KEXT	TOOL
P1	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Entre 2 y 4 repositorios con tecnología no compatible para la integración (3).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P2	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	No se utilizan tablas auxiliares (1).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
P3	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P4	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (4).	Entre 2 y 4 repositorios con tecnología no compatible para la integración (3).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P5	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos (1).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
P6	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 101 y 1.000 tuplas en la tabla principal (2).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).

Tabla B.7. Caracterización de los proyectos para el modelo de estimación de esfuerzo (P1 a P6).

#	OBTY	LECO	AREP	QTUM	QTUA	KLDS	KEXT	TOOL
P7	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 20.001 y 80.000 tuplas en la tabla principal (4).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P8	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (4).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos (1).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P9	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre la identificación de una clase desconocida previamente (5).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
P10	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 101 y 1.000 tuplas en la tabla principal (2).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P11	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Sólo 1 repositorio disponible (1).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).

Tabla B.8. Caracterización de los proyectos para el modelo de estimación de esfuerzo (P7 a P11).

#	OBTY	LECO	AREP	QTUM	QTUA	KLDS	KEXT	TOOL
P12	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos (3).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
P13	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Hasta 100 tuplas en la tabla principal (1).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Las tablas y repositorios no están documentados y existen expertos en los datos pero no están disponibles para explicarlos (5).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P14	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P15	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Sólo 1 repositorio disponible (1).	Entre 101 y 1.000 tuplas en la tabla principal (2).	No se utilizan tablas auxiliares (1).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P16	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (4).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos (1).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).

Tabla B.9. Caracterización de los proyectos para el modelo de estimación de esfuerzo (P12 a P16).

#	OBTY	LECO	AREP	QTUM	QTUA	KLDS	KEXT	TOOL
P17	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Sólo 1 repositorio disponible (1).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P18	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P19	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
P20	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre la identificación de una clase desconocida previamente (5).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar (4).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Las tablas y repositorios no están documentados y existen expertos en los datos pero no están disponibles para explicarlos (5).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P21	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Sólo 1 repositorio disponible (1).	Hasta 100 tuplas en la tabla principal (1).	No se utilizan tablas auxiliares (1).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
P22	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	Más de 50.000 tuplas en las tablas auxiliares (4).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).

Tabla B.10. Caracterización de los proyectos para el modelo de estimación de esfuerzo (P17 a P22).

#	OBTY	LECO	AREP	QTUM	QTUA	KLDS	KEXT	TOOL
P23	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida (4).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos (3).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de dato (5).
P24	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida. (4).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente. (3).	Sólo 1 repositorio disponible (1).	Entre 101 y 1.000 tuplas en la tabla principal (2).	No se utilizan tablas auxiliares (1).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
P25	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente. (3).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente. (3).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Hasta 100 tuplas en la tabla principal (1).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Las tablas y repositorios no están documentados y existen expertos en los datos pero no están disponibles para explicarlos (5).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P26	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos (2).	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos (3).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
P27	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida. (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto (1).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 20.001 y 80.000 tuplas en la tabla principal (4).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos (3).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
P28	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Sólo 1 repositorio disponible (1).	Entre 101 y 1.000 tuplas en la tabla principal (2).	Hasta 1.000 tuplas en las tablas auxiliares (2).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos (5).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).

Tabla B.11. Caracterización de los proyectos para el modelo de estimación de esfuerzo (P23 a P28).

#	OBTY	LECO	AREP	QTUM	QTUA	KLDS	KEXT	TOOL
P29	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 1.001 y 20.000 tuplas en la tabla principal (3).	Entre 1.001 y 50.000 tuplas en las tablas auxiliares (3).	Las tablas y repositorios no están documentados y existen expertos en los datos pero no están disponibles para explicarlos (5).	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos (3).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, y permite importar más de una tabla de datos en forma independiente (2).
P30	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre la identificación de una clase desconocida previamente (5).	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente (3).	Entre 2 y 4 repositorios con tecnología compatible para la integración (2).	Entre 80.001 y 5.000.000 tuplas en la tabla principal (5).	Más de 50.000 tuplas en las tablas auxiliares (4).	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos (4).	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos (4).	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos (5).
P31	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida (1).	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto. (1).	Sólo 1 repositorio disponible (1).	Entre 101 y 1.000 tuplas en la tabla principal (2).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos (2).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos (3).
P32	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente (3).	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto (2).	Sólo 1 repositorio disponible (1).	Entre 101 y 1.000 tuplas en la tabla principal (2).	No se utilizan tablas auxiliares (1).	Todas las tablas y repositorios están correctamente documentados (1).	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos (1).	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, y permite importar más de una tabla de datos en forma independiente (2).

Tabla B.12. Caracterización de los proyectos para el modelo de estimación de esfuerzo (P29 a P32).

ANEXO C: APLICACIÓN DEL MODELO PARA LA EVALUACIÓN DE VIABILIDAD EN PROYECTOS DE EXPLOTACIÓN PARA SU VALIDACIÓN

En este anexo se indican los resultados de aplicar el modelo para la Evaluación de la Viabilidad en los treinta y siete proyectos de explotación de información que fueron presentados en el Anexo B de este trabajo de tesis.

A partir de la caracterización de los proyectos (que fue indicada en la Tabla B.5 del Anexo B), se aplican los cinco pasos del modelo propuesto. Como resultado se generan los intervalos difusos de cada dimensión (I_d) con su correspondiente representación gráfica, el valor numérico por dimensión (V_d) y el valor global de la viabilidad del proyecto (EV).

Para cada proyecto se muestran los resultados obtenidos en su tabla correspondiente (Tablas C.1 a C.37). Nótese que, en cada caso, la representación gráfica se muestra también el umbral mínimo de referencia (es decir, el intervalo correspondiente al valor 'regular') con una tonalidad gris más clara.

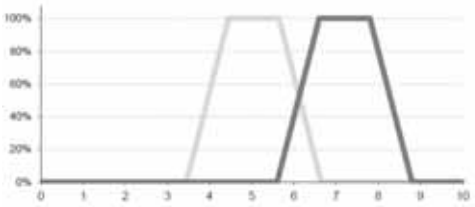
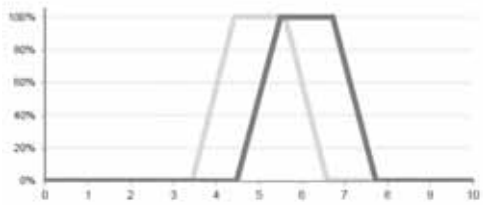
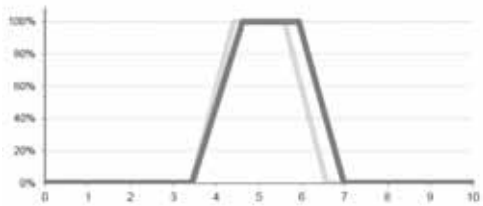
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P1	*Plausibilidad*	*(5,6; 6,6; 7,8; 8,8)* 	*7,20*	*Dimensión aceptable que supera en forma holgada el umbral mínimo con un valor equivalente a 'mucho'*
	Adecuación	*(4,5; 5,5; 6,7; 7,7)* 	*6,11*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(3,4; 4,6; 5,9; 7,0)* 	*5,25*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Valor global de la viabilidad del proyecto (*EV*)		***6,27***	***VIABLE***

Tabla C.1. Resultados de aplicar el modelo al proyecto P1.

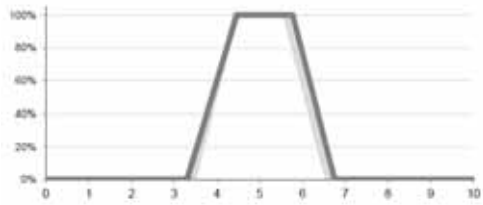
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P2	*Plausibilidad*	*(5,3; 6,3; 7,5; 8,3)* 	*6,87*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(3,3; 4,5; 5,8; 6,8)* 	*5,07*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Éxito	*(3,4; 4,6; 5,9; 7)* 	*5,25*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Valor global de la viabilidad del proyecto (*EV*)		***5,77***	***VIABLE***

Tabla C.2. Resultados de aplicar el modelo al proyecto P2.

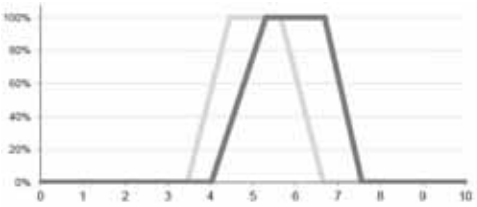
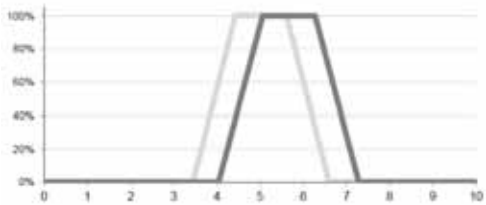
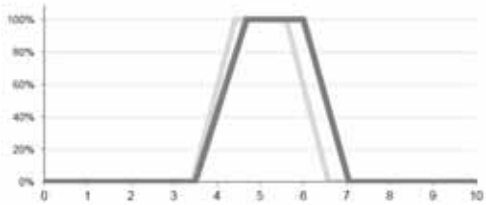
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P3	*Plausibilidad*	*(4; 5,3; 6,7; 7,6)* 	*5,90*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Adecuación	*(4; 5,1; 6,3; 7,3)* 	*5,67*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Éxito	*(3,5; 4,7; 6; 7,1)* 	*5,31*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Valor global de la viabilidad del proyecto (*EV*)		***5,65***	***VIABLE***

Tabla C.3. Resultados de aplicar el modelo al proyecto P3.

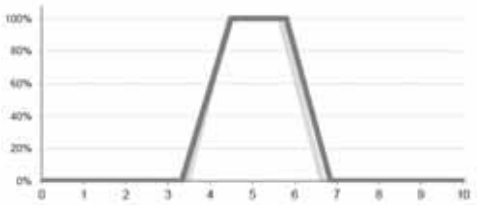
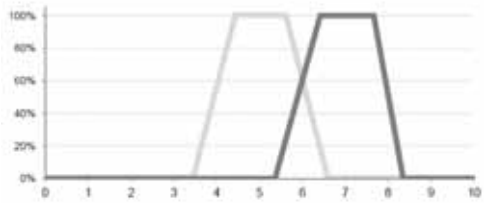
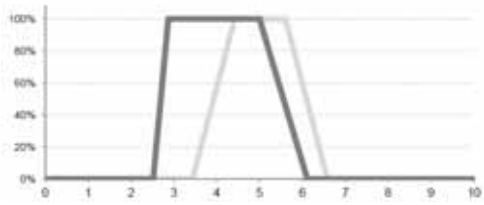
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P4	*Plausibilidad*	*(3,3; 4,5; 5,8; 6,9)* 	*5,12*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Adecuación	*(5,4; 6,4; 7,7; 8,4)* 	*6,95*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor cercano a 'mucho'*
	Éxito	*(2,5; 2,9; 5; 6,1)* 	*4,12*	*Dimensión NO aceptable pero con un valor tendiendo al umbral mínimo*
	Valor global de la viabilidad del proyecto (*EV*)		***5,51***	***NO VIABLE sólo por dimensión***

Tabla C.4. Resultados de aplicar el modelo al proyecto P4.

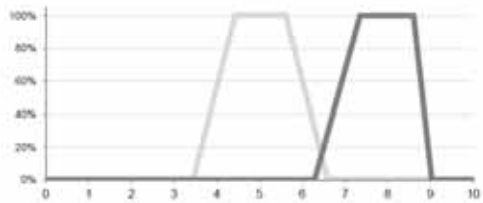
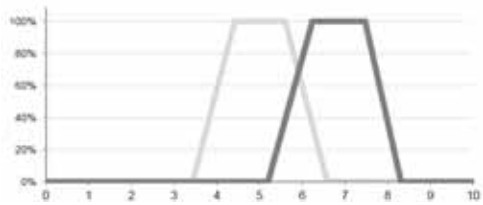
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P5	*Plausibilidad*	*(3,3; 4,5; 5,8; 6,9)* 	*5,12*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Adecuación	*(6,3; 7,4; 8,6; 9)* 	*7,82*	*Dimensión aceptable que supera en forma muy holgada el umbral mínimo con un valor mayor a 'mucho'*
	Éxito	*(5,2; 6,2; 7,5; 8,3)* 	*6,81*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Valor global de la viabilidad del proyecto (*EV*)		***6,56***	***VIABLE***

Tabla C.5. Resultados de aplicar el modelo al proyecto P5.

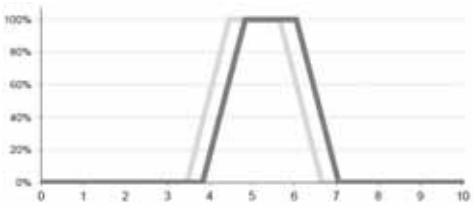
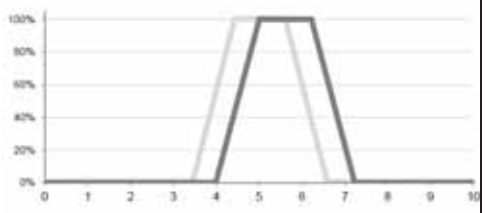
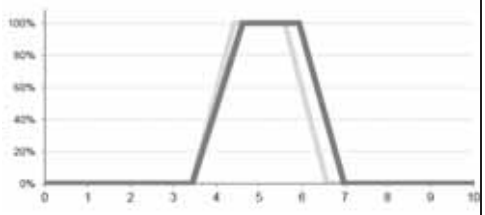
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P6	*Plausibilidad*	*(3,8; 4,8; 6,1; 7,1)* 	*5,45*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Adecuación	*(4; 5; 6,2; 7,2)* 	*5,61*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Éxito	*(3,4; 4,6; 5,9; 7)* 	*5,25*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Valor global de la viabilidad del proyecto (*EV*)		***5,45***	***VIABLE***

Tabla C.6. Resultados de aplicar el modelo al proyecto P6.

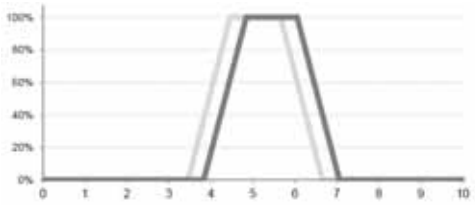
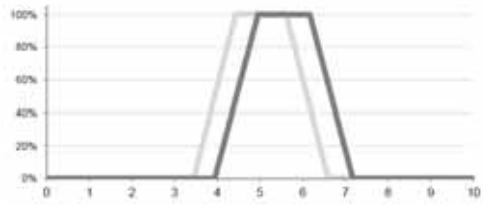
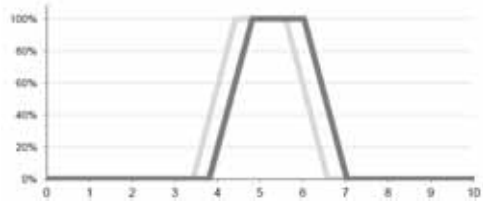
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P7	*Plausibilidad*	*(3,8; 4,8; 6,1; 7,1)* 	*5,45*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Adecuación	*(3,9; 5; 6,2; 7,2)* 	*5,56*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Éxito	*(3,8; 4,8; 6; 7)* 	*5,42*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***5,48***	***VIABLE***

Tabla C.7. Resultados de aplicar el modelo al proyecto P7.

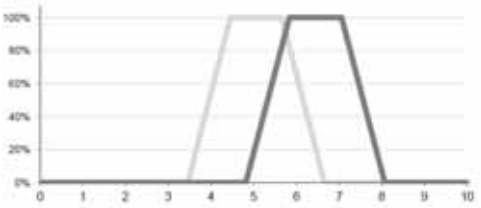
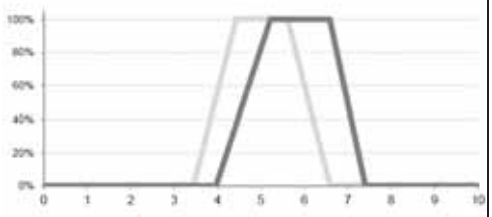
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P8	*Plausibilidad*	*(4,8; 5,8; 7,1; 8,1)* 	*6,45*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(4; 5,2; 6,6; 7,4)* 	*5,80*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Éxito	*(3,4; 4,6; 5,9; 6,8)* 	*5,18*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Valor global de la viabilidad del proyecto (*EV*)		***5,87***	***VIABLE***

Tabla C.8. Resultados de aplicar el modelo al proyecto P8.

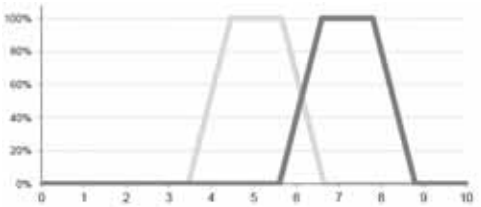
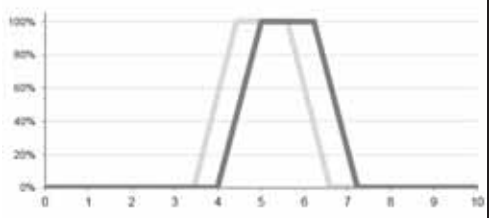
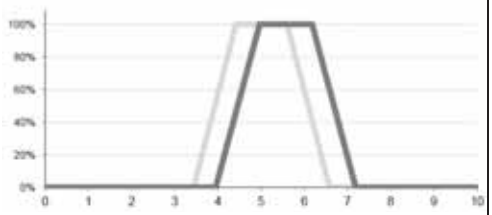
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P9	*Plausibilidad*	*(5,6; 6,6; 7,8; 8,8)* 	*7,20*	*Dimensión aceptable que supera en forma holgada el umbral mínimo con un valor equivalente a 'mucho'*
	Adecuación	*(4; 5; 6,2; 7,2)* 	*5,61*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Éxito	*(3,9; 5; 6,2; 7,2)* 	*5,57*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***6,18***	***VIABLE***

Tabla C.9. Resultados de aplicar el modelo al proyecto P9.

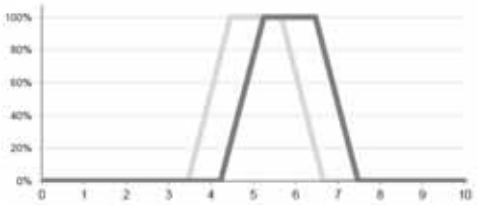
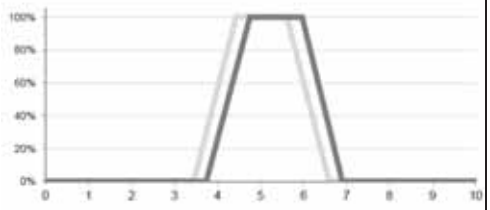
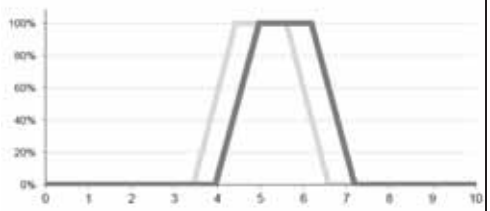
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P10	*Plausibilidad*	*(4,2; 5,2; 6,5; 7,5)* 	*5,85*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Adecuación	*(3,7; 4,8; 6; 6,9)* 	*5,34*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Éxito	*(3,9; 5; 6,2; 7,2)* 	*5,57*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***5,59***	***VIABLE***

Tabla C.10. Resultados de aplicar el modelo al proyecto P10.

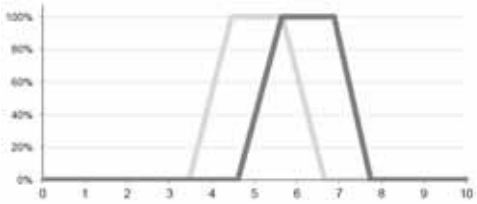
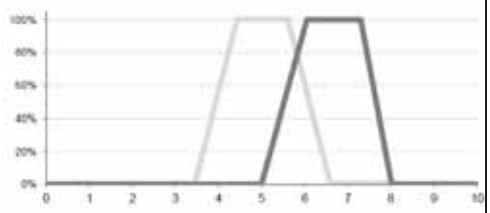
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P11	*Plausibilidad*	*(4,6; 5,6; 6,9; 7,7)* 	*6,22*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(5; 6,1; 7,3; 8)* 	*6,59*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(3,8; 4,8; 6; 7)* 	*5,42*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***6,14***	***VIABLE***

Tabla C.11. Resultados de aplicar el modelo al proyecto P11.

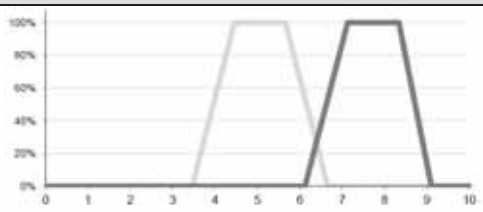
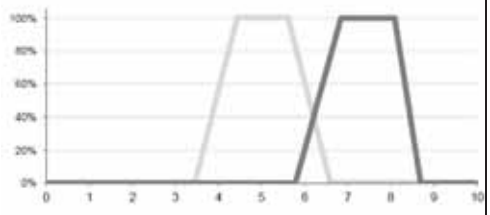
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P12	*Plausibilidad*	*(6,1; 7,1; 8,3; 9,1)* 	*7,67*	*Dimensión aceptable que supera en forma holgada el umbral mínimo con un valor mayor a 'mucho'*
	Adecuación	*(5,8; 6,8; 8,1; 8,7)* 	*7,35*	*Dimensión aceptable que supera en forma holgada el umbral mínimo con un valor equivalente a 'mucho'*
	Éxito	*(4,8; 5,8; 7,1; 8,1)* 	*6,45*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Valor global de la viabilidad del proyecto (*EV*)		***7,22***	***VIABLE***

Tabla C.12. Resultados de aplicar el modelo al proyecto P12.

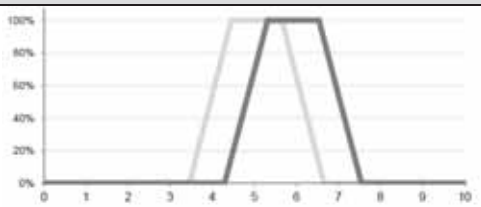
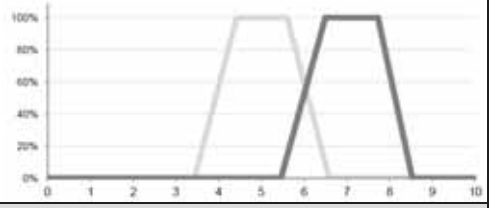
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P13	*Plausibilidad*	*(4,3; 5,3; 6,5; 7,5)* 	*5,93*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Adecuación	*(3,3; 4,5; 5,8; 6,8)* 	*5,09*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Éxito	*(5,5; 6,5; 7,7; 8,5)* 	*7,05*	*Dimensión aceptable que supera en forma holgada el umbral mínimo con un valor cercano a 'mucho'*
	Valor global de la viabilidad del proyecto (*EV*)		***5,93***	***VIABLE***

Tabla C.13. Resultados de aplicar el modelo al proyecto P13.

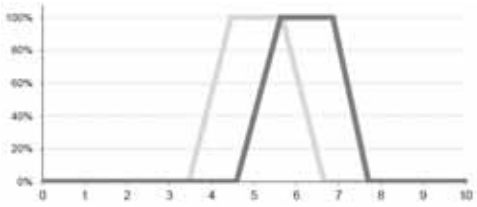
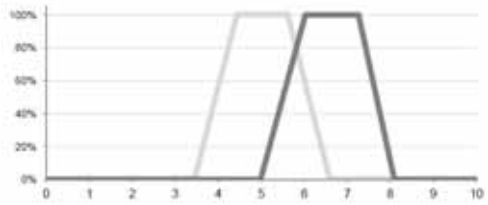
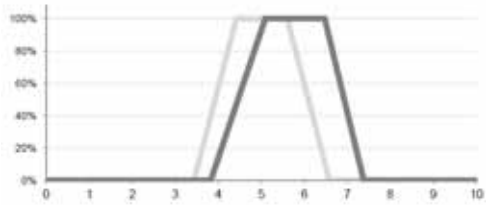
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P14	*Plausibilidad*	*(4,6; 5,6; 6,9; 7,7)* 	*6,20*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(5; 6; 7,3; 8,1)* 	*6,59*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(3,8; 5,1; 6,5; 7,4)* 	*5,69*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***6,20***	***VIABLE***

Tabla C.14. Resultados de aplicar el modelo al proyecto P14.

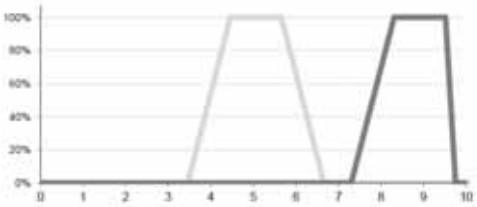
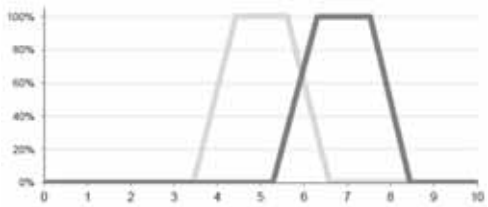
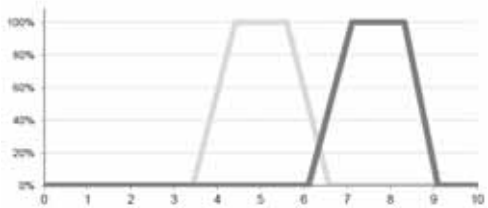
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P15	*Plausibilidad*	*(7,3; 8,3; 9,5; 9,7)* 	*8,72*	*Dimensión aceptable que supera en forma muy holgada el umbral mínimo con un valor tendiendo a 'todo'*
	Adecuación	*(5,3; 6,3; 7,5; 8,5)* 	*6,89*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(6,1; 7,1; 8,3; 9,1)* 	*7,66*	*Dimensión aceptable que supera en forma holgada el umbral mínimo con un valor mayor a 'mucho'*
	Valor global de la viabilidad del proyecto (*EV*)		***7,77***	***VIABLE***

Tabla C.15. Resultados de aplicar el modelo al proyecto P15.

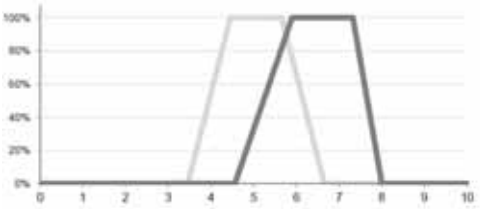
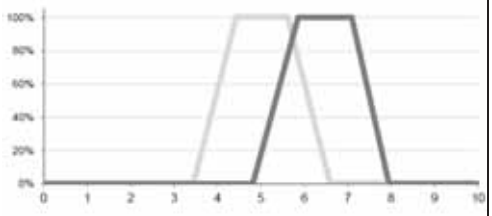
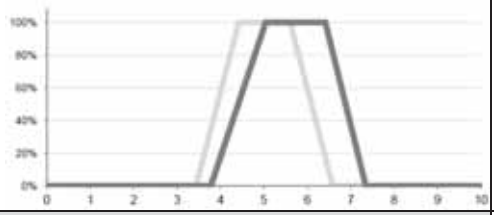
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P16	*Plausibilidad*	*(4,6; 5,9; 7,3; 8)* 	*6,45*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(4,8; 5,9; 7,1; 7,9)* 	*6,43*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(3,8; 5; 6,4; 7,3)* 	*5,64*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***6,22***	***VIABLE***

Tabla C.16. Resultados de aplicar el modelo al proyecto P16.

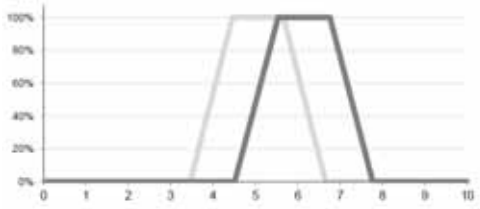
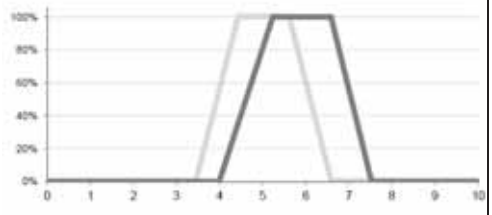
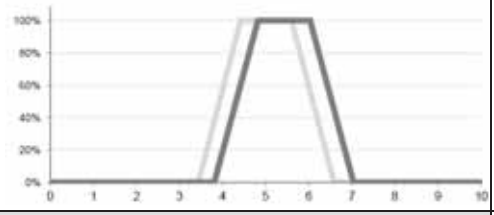
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P17	*Plausibilidad*	*(4,5; 5,5; 6,8; 7,8)* 	*6,14*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(4; 5,2; 6,6; 7,5)* 	*5,83*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Éxito	*(3,8; 4,8; 6; 7)* 	*5,42*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***5,83***	***VIABLE***

Tabla C.17. Resultados de aplicar el modelo al proyecto P17.

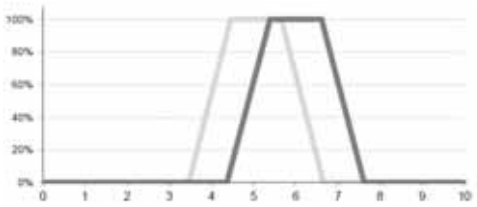
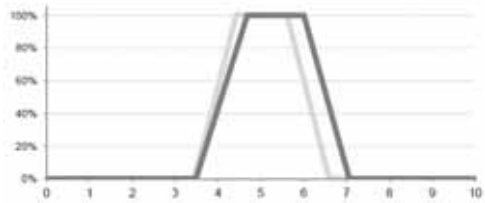
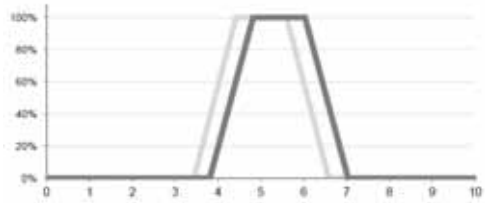
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P18	*Plausibilidad*	*(4,4; 5,4; 6,6; 7,6)* 	*6,00*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(3,5; 4,7; 6; 7,1)* 	*5,31*	*Dimensión ac0eptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Éxito	*(3,8; 4,8; 6; 7)* 	*5,42*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***5,59***	***VIABLE***

Tabla C.18. Resultados de aplicar el modelo al proyecto P18.

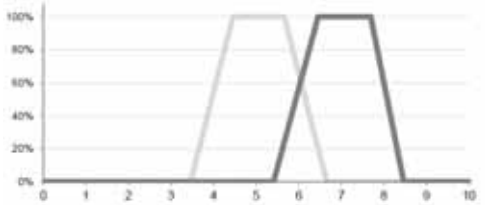
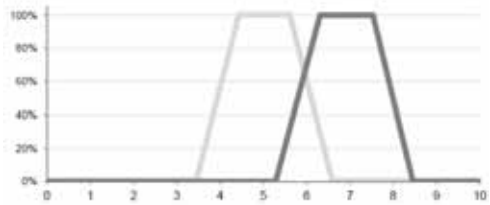
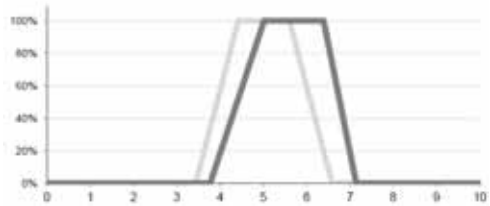
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P19	*Plausibilidad*	*(5,4; 6,5; 7,7; 8,5)* 	*7,01*	*Dimensión aceptable que supera en forma holgada el umbral mínimo con un valor cercano a 'mucho'*
	Adecuación	*(5,3; 6,3; 7,5; 8,5)* 	*6,89*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(3,8; 5; 6,4; 7,1)* 	*5,58*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***6,58***	***VIABLE***

Tabla C.19. Resultados de aplicar el modelo al proyecto P19.

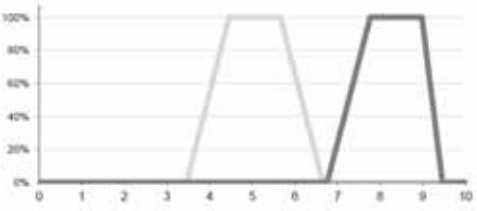
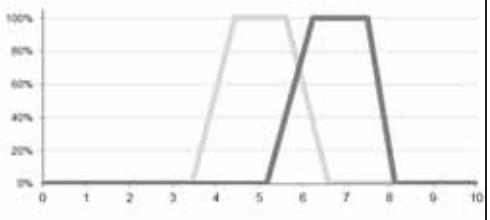
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P20	*Plausibilidad*	*(6,8; 7,8; 9; 9,5)* 	*8,24*	*Dimensión aceptable que supera en forma muy holgada el umbral mínimo con un valor tendiendo a 'todo'*
	Adecuación	*(5,2; 6,2; 7,5; 8,1)* 	*6,75*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(3,7; 4,9; 6,3; 7,2)* 	*5,52*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***6,96***	***VIABLE***

Tabla C.20. Resultados de aplicar el modelo al proyecto P20.

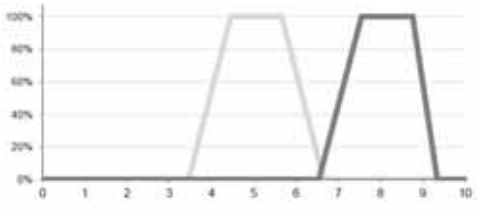
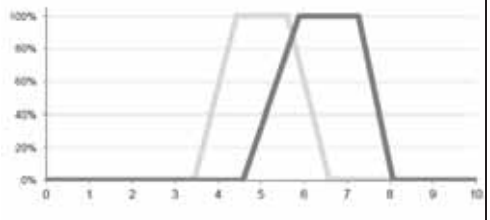
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P21	*Plausibilidad*	*(6,5; 7,5; 8,8; 9,3)* 	*8,05*	*Dimensión aceptable que supera en forma muy holgada el umbral mínimo con un valor tendiendo a 'todo'*
	Adecuación	*(4,6; 5,9; 7,3; 8,1)* 	*6,45*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(3,4; 4,7; 6; 6,9)* 	*5,25*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Valor global de la viabilidad del proyecto (*EV*)		***6,70***	***VIABLE***

Tabla C.21. Resultados de aplicar el modelo al proyecto P21.

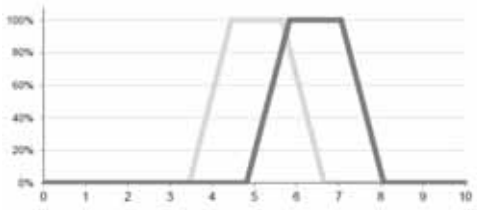
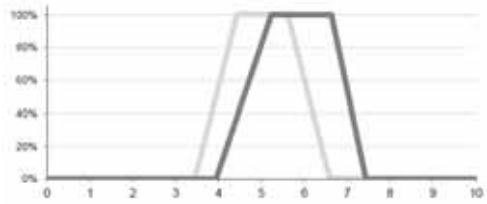
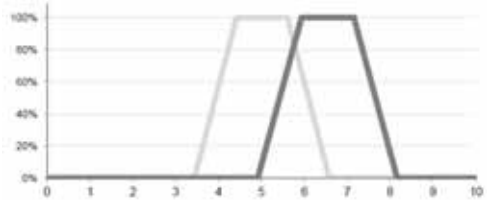
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P22	*Plausibilidad*	*(4,8; 5,8; 7,1; 8,1)* 	*6,45*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(3,9; 5,2; 6,6; 7,4)* 	*5,81*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Éxito	*(4,9; 5,9; 7,2; 8,2)* 	*6,54*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Valor global de la viabilidad del proyecto (*EV*)		***6,24***	***VIABLE***

Tabla C.22. Resultados de aplicar el modelo al proyecto P22.

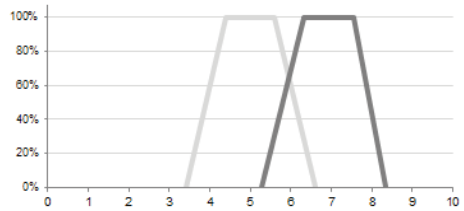
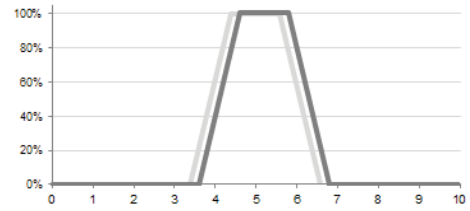
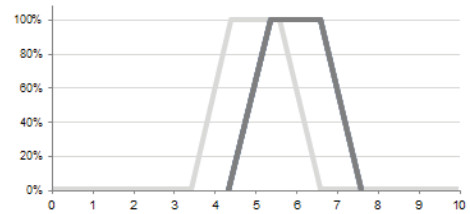
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P23	*Plausibilidad*	*(5,3; 6,3; 7,5; 8,3)* 	*6,87*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(3,6; 4,6; 5,8; 6,8)* 	*5,20*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Éxito	*(4,3; 5,4; 6,6; 7,6)* 	*5,96*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***6,01***	***VIABLE***

Tabla C.23. Resultados de aplicar el modelo al proyecto P23.

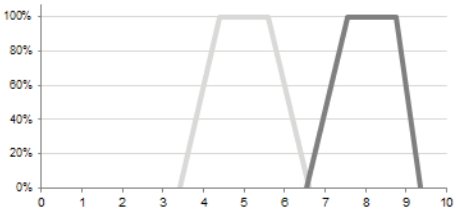
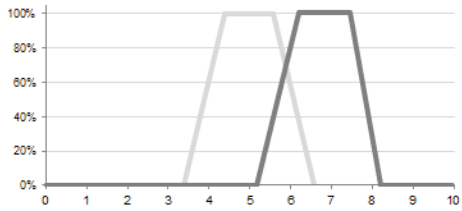
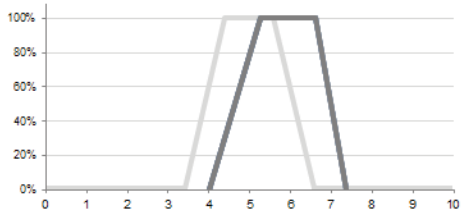
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P24	*Plausibilidad*	*(6,5; 7,5; 8,8; 9,3)* 	*8,05*	*Dimensión aceptable que supera muy bien el umbral mínimo con un valor tendiendo a 'todo'*
	Adecuación	*(5,2; 6,2; 7,5; 8,2)* 	*6,76*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(4; 5,3; 6,6; 7,4)* 	*5,81*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***6,97***	***VIABLE***

Tabla C.24. Resultados de aplicar el modelo al proyecto P24.

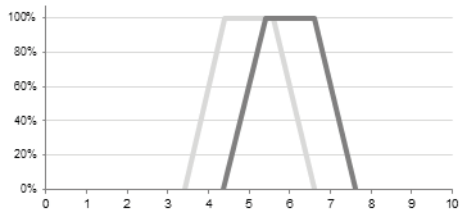
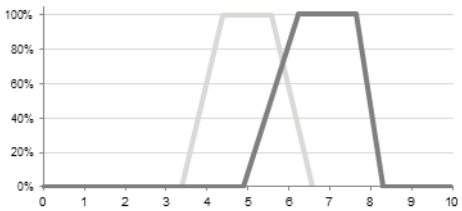
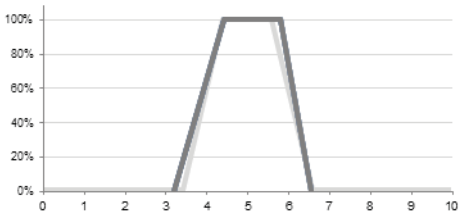
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P25	*Plausibilidad*	*(4,4; 5,4; 6,6; 7,6)* 	*6,00*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(4,9; 6,2; 7,6; 8,3)* 	*6,76*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(3,2; 4,4; 5,8; 6,6)* 	*5,00*	*Dimensión aceptable que se encuentra justo en el umbral mínimo por lo que debe ser muy monitoreado y controlado*
	Valor global de la viabilidad del proyecto (*EV*)		***6,00***	***VIABLE***

Tabla C.25. Resultados de aplicar el modelo al proyecto P25.

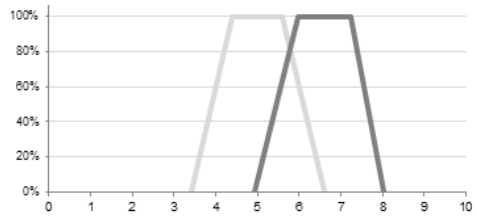
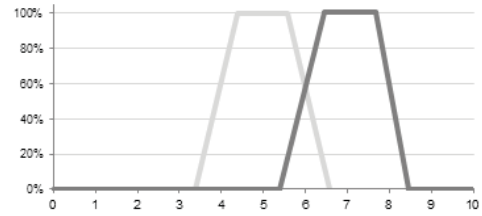
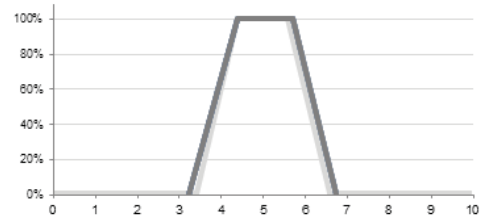
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P26	*Plausibilidad*	*(4,9; 6; 7,2; 8)* 	*6,55*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(5,4; 6,4; 7,7; 8,5)* 	*7,00*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(3,2; 4,4; 5,7; 6,7)* 	*5,01*	*Dimensión aceptable que se encuentra justo en el umbral mínimo por lo que debe ser muy monitoreado y controlado*
	Valor global de la viabilidad del proyecto (*EV*)		***6,29***	***VIABLE***

Tabla C.26. Resultados de aplicar el modelo al proyecto P26.

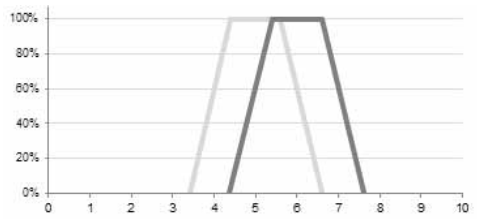
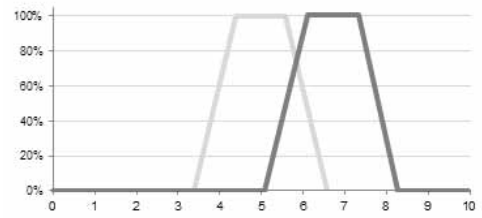
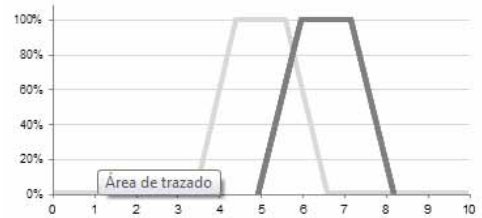
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P27	*Plausibilidad*	*(4,4; 5,4; 6,6; 7,6)* 	*6,00*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(5,1; 6,1; 7,3; 8,3)* 	*6,70*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(4,9; 5,9; 7,2; 8,2)* 	*6,54*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Valor global de la viabilidad del proyecto (*EV*)		***6,40***	***VIABLE***

Tabla C.27. Resultados de aplicar el modelo al proyecto P27.

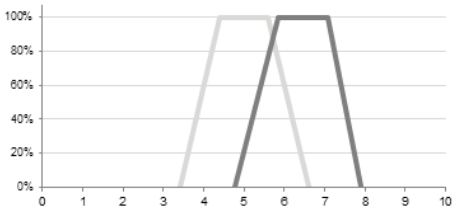
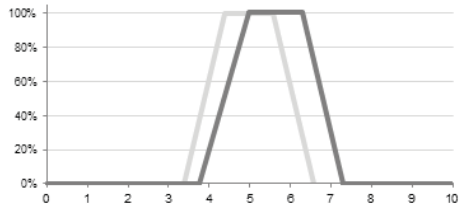
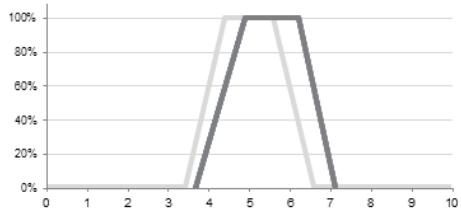
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P28	*Plausibilidad*	*(4,8; 5,8; 7,1; 7,9)* 	*6,39*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Adecuación	*(3,8; 5; 6,3; 7,3)* 	*5,58*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Éxito	*(3,7; 4,9; 6,2; 7,1)* 	*5,47*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Valor global de la viabilidad del proyecto (*EV*)		***5,85***	***VIABLE***

Tabla C.28. Resultados de aplicar el modelo al proyecto P28.

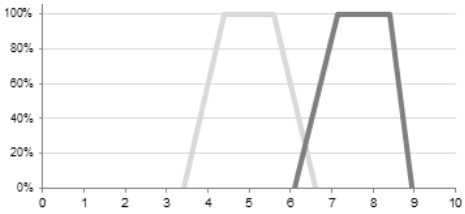
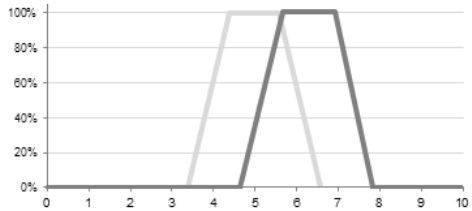
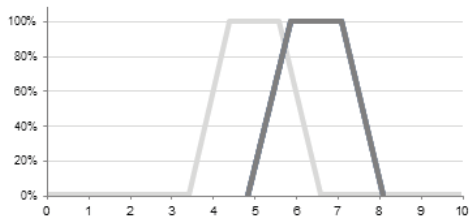
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P29	*Plausibilidad*	*(6,1; 7,1; 8,4; 8,9)* 	*7,64*	*Dimensión aceptable que supera muy bien el umbral mínimo con un valor tendiendo a 'todo'*
	Adecuación	*(4,6; 5,7; 6,9; 7,8)* 	*6,27*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Éxito	*(4,8; 5,8; 7,1; 8,1)* 	*6,45*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor tendiendo a 'mucho'*
	Valor global de la viabilidad del proyecto (*EV*)		***6,82***	***VIABLE***

Tabla C.29. Resultados de aplicar el modelo al proyecto P29.

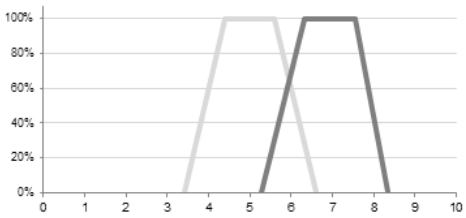
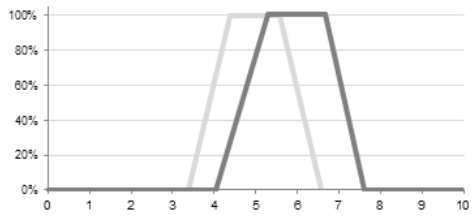
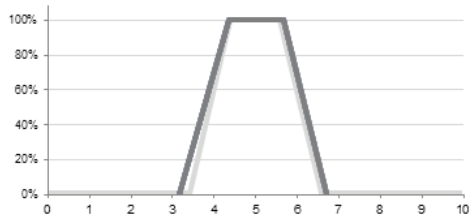
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P30	*Plausibilidad*	*(5,3; 6,3; 7,5; 8,3)* 	*6,87*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor cercano a 'mucho'*
	Adecuación	*(4; 5,3; 6,7; 7,6)* 	*5,90*	*Dimensión aceptable pero se encuentra cercano al umbral mínimo por lo que debe ser monitoreado*
	Éxito	*(3,2; 4,3; 5,7; 6,7)* 	*4,97*	*Dimensión NO aceptable pero con un valor muy cercano al umbral mínimo*
	Valor global de la viabilidad del proyecto (*EV*)		***6,00***	***NO VIABLE sólo por dimensión***

Tabla C.30. Resultados de aplicar el modelo al proyecto P30.

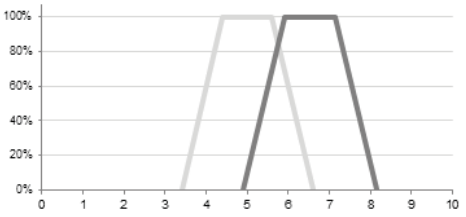
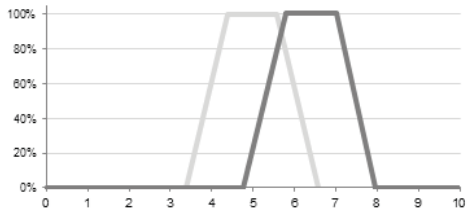
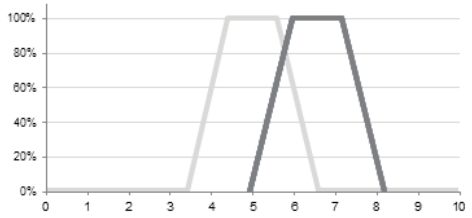
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P31	*Plausibilidad*	*(4,9; 5,9; 7,1; 8,1)* 	*6,52*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor cercano a 'mucho'*
	Adecuación	*(4,8; 5,8; 7; 8)* 	*6,39*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor cercano a 'mucho'*
	Éxito	*(4,9; 5,9; 7,2; 8,2)* 	*6,54*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor cercano a 'mucho'*
	Valor global de la viabilidad del proyecto (*EV*)		***6,48***	***VIABLE***

Tabla C.31. Resultados de aplicar el modelo al proyecto P31.

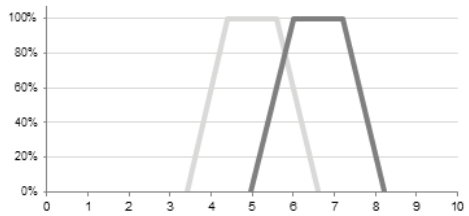
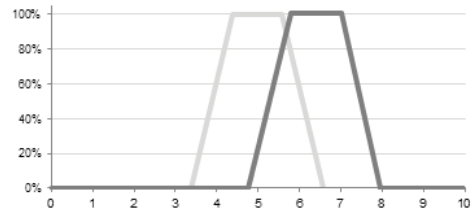
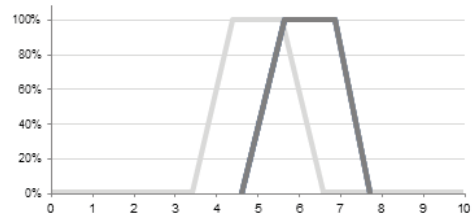
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P32	*Plausibilidad*	*(5; 6; 7,2; 8,2)* 	*6,60*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor cercano a 'mucho'*
	Adecuación	*(4,8; 5,8; 7; 8)* 	*6,39*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor cercano a 'mucho'*
	Éxito	*(4,6; 5,6; 6,9; 7,7)* 	*6,20*	*Dimensión aceptable que supera en forma razonable el umbral mínimo con un valor cercano a 'mucho'*
	Valor global de la viabilidad del proyecto (*EV*)		***6,42***	***VIABLE***

Tabla C.32. Resultados de aplicar el modelo al proyecto P32.

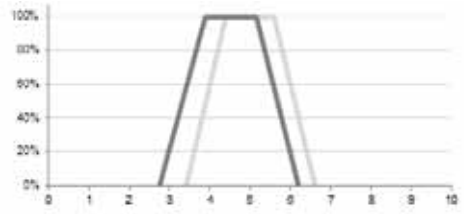
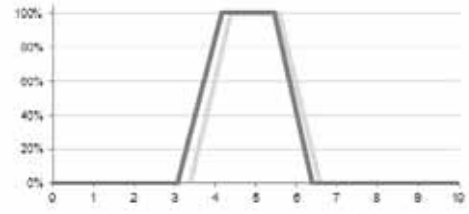
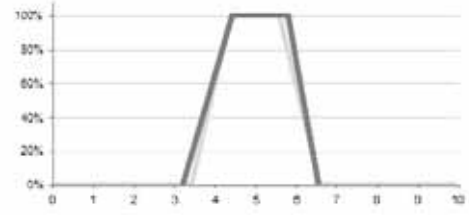
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P33	*Plausibilidad*	*(2,7; 3,9; 5,2; 6,2)* 	*4,49*	*Dimensión NO aceptable pero con un valor cercano al umbral mínimo*
	Adecuación	*(3,1; 4,2; 5,4; 6,4)* 	*4,77*	*Dimensión NO aceptable pero con un valor cercano al umbral mínimo*
	Éxito	*(3,2; 4,4; 5,8; 6,6)* 	*4,99*	*Dimensión NO aceptable pero con un valor muy cercano al umbral mínimo*
	Valor global de la viabilidad del proyecto (*EV*)		***4,73***	***NO VIABLE***

Tabla C.33. Resultados de aplicar el modelo al proyecto P33.

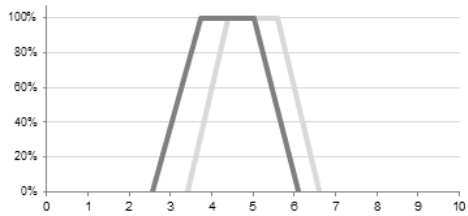
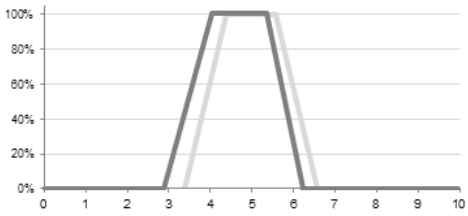
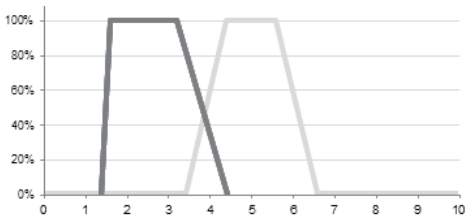
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P34	*Plausibilidad*	*(2,6; 3,7; 5; 6,1)* 	*4,36*	*Dimensión NO aceptable pero con un valor cercano al umbral mínimo*
	Adecuación	*(2,9; 4; 5,4; 6,2)* 	*4,62*	*Dimensión NO aceptable pero con un valor cercano al umbral mínimo*
	Éxito	*(1,4; 1,6; 3,2; 4,4)* 	*2,64*	*Dimensión NO aceptable y con un valor tendiendo a 'poco'*
	Valor global de la viabilidad del proyecto (*EV*)		***3,99***	***NO VIABLE***

Tabla C.34. Resultados de aplicar el modelo al proyecto P34.

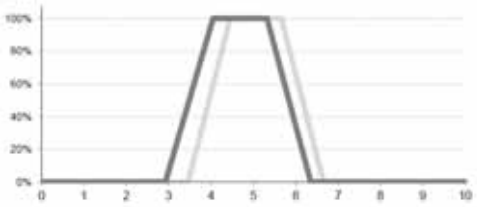
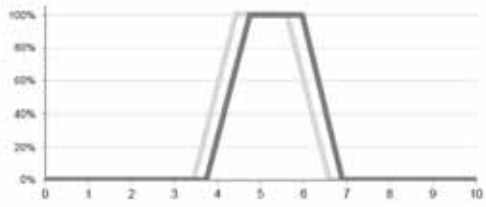
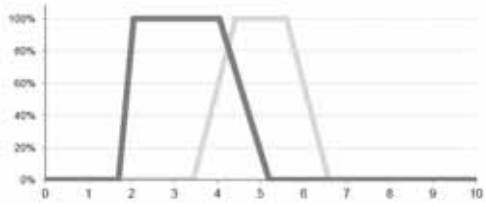
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P35	*Plausibilidad*	*(2,9; 4; 5,3; 6,4)* 	*4,66*	*Dimensión NO aceptable pero con un valor cercano al umbral mínimo*
	Adecuación	*(3,7; 4,8; 6; 6,9)* 	*5,34*	*Dimensión aceptable pero se encuentra muy cercano al umbral mínimo por lo que debe ser monitoreado rigurosamente*
	Éxito	*(1,7; 2,1; 4; 5,2)* 	*3,25*	*Dimensión NO aceptable y con un valor tendiendo a 'poco'*
	Valor global de la viabilidad del proyecto (*EV*)		***4,52***	***NO VIABLE***

Tabla C.35. Resultados de aplicar el modelo al proyecto P35.

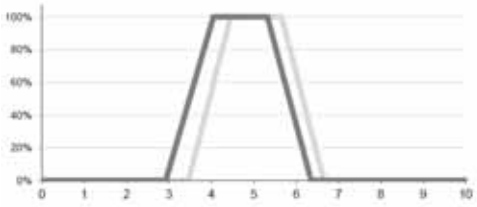
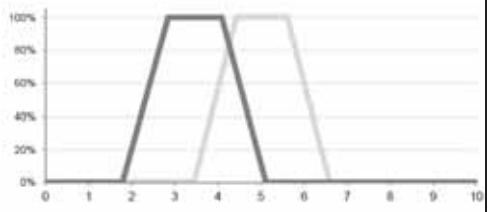
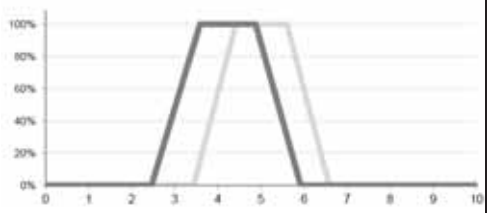
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P36	*Plausibilidad*	*(2,9; 4; 5,3; 6,4)* 	*4,66*	*Dimensión NO aceptable pero con un valor cercano al umbral mínimo*
	Adecuación	*(1,8; 2,8; 4,1; 5,1)* 	*3,46*	*Dimensión NO aceptable y con un valor tendiendo a 'poco'*
	Éxito	*(2,5; 3,6; 4,9; 5,9)* 	*4,21*	*Dimensión NO aceptable pero con un valor tendiendo al umbral mínimo*
	Valor global de la viabilidad del proyecto (*EV*)		***4,10***	***NO VIABLE***

Tabla C.36. Resultados de aplicar el modelo al proyecto P36.

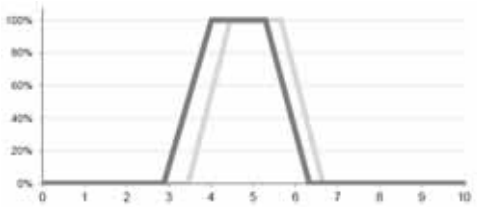
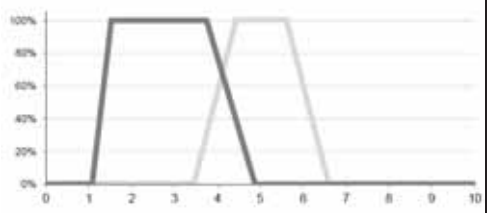
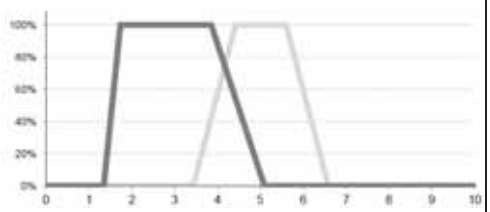
#	Dimensión	Intervalo Valor de la Dimensión (I_d)	Valor de la Dimensión (V_d)	Interpretación
P37	*Plausibilidad*	*(2,9; 4; 5,3; 6,3)* 	*4,63*	*Dimensión NO aceptable pero con un valor cercano al umbral mínimo*
	Adecuación	*(1,1; 1,5; 3,7; 4,9)* 	*2,81*	*Dimensión NO aceptable y con un valor equivalente a 'poco'*
	Éxito	*(1,3; 1,7; 3,9; 5,1)* 	*3,01*	*Dimensión NO aceptable y con un valor tendiendo a 'poco'*
	Valor global de la viabilidad del proyecto (*EV*)		***3,52***	***NO VIABLE***

Tabla C.37. Resultados de aplicar el modelo al proyecto P37.

ANEXO D: APLICACIÓN DEL MODELO PARA LA ESTIMACIÓN DE ESFUERZO EN PROYECTOS DE EXPLOTACIÓN PARA SU VALIDACIÓN

En este anexo se indican los resultados de aplicar el modelo para la Estimación de Esfuerzo en los primeros treinta y dos proyectos de explotación de información que fueron presentados en el Anexo B de este trabajo de tesis.

A partir de la caracterización de los proyectos (indicada en las Tablas B.6 a B.11 del Anexo B), se aplican los dos métodos de estimación propuestos. Los resultados obtenidos se muestran en las Tablas D.1 a D.32 indicando en cada una el esfuerzo estimado por la fórmula lineal (PEM_L) y el esfuerzo estimado por el método empírico (PEM_E).

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P1</td><td>Fórmula Lineal</td><td colspan="4">PEM_L = (0,80 . 1) + (1,10 . 1) - (1,20 . 3) - (0,30 . 3) - (0,70 . 1) + (1,80 . 3) - (0,90 . 2) + (1,86 . 3) - 3,30</td><td>2,58 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">3,57 meses/hombre</td></tr>
<tr><td>2,00</td><td>2,40</td><td>0,60</td><td>0,70</td></tr>
<tr><td colspan="4">PEM_E = [(1,80 . 2,00) + (0,90 . 2,40) + (1,40 . 0,60) - 1,50] . 0,70</td></tr>
</table>

Tabla D.1. Resultados de aplicar los métodos del modelo al proyecto P1.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P2</td><td>Fórmula Lineal</td><td colspan="4">PEM_L = (0,80 . 1) + (1,10 . 1) - (1,20 . 1) - (0,30 . 3) - (0,70 . 1) + (1,80 . 3) - (0,90 . 5) + (1,86 . 5) - 3,30</td><td>6,00 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">7,23 meses/hombre</td></tr>
<tr><td>2,00</td><td>1,50</td><td>2,70</td><td>1,00</td></tr>
<tr><td colspan="4">PEM_E = [(1,80 . 2,00) + (0,90 . 1,50) + (1,40 . 2,70) - 1,50] . 1,00</td></tr>
</table>

Tabla D.2. Resultados de aplicar los métodos del modelo al proyecto P2.

#	Método	Aplicación del Método				Esfuerzo Estimado
P3	*Fórmula Lineal*	$PEM_L = (0{,}80 \cdot 4) + (1{,}10 \cdot 1) - (1{,}20 \cdot 1) - (0{,}30 \cdot 3) - (0{,}70 \cdot 3) + (1{,}80 \cdot 2) - (0{,}90 \cdot 5) + (1{,}86 \cdot 3) - 3{,}30$				***1,48 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***3,58 meses/hombre***
		1,00	*2,40*	*0,80*	*1,00*	
		$PEM_E = [(1{,}80 \cdot 1{,}00) + (0{,}90 \cdot 2{,}40) + (1{,}40 \cdot 0{,}80) - 1{,}50] \cdot 1{,}00$				

Tabla D.3. Resultados de aplicar los métodos del modelo al proyecto P3.

#	Método	Aplicación del Método				Esfuerzo Estimado
P4	*Fórmula Lineal*	$PEM_L = (0{,}80 \cdot 1) + (1{,}10 \cdot 4) - (1{,}20 \cdot 3) - (0{,}30 \cdot 5) - (0{,}70 \cdot 1) + (1{,}80 \cdot 1) - (0{,}90 \cdot 2) + (1{,}86 \cdot 3) - 3{,}30$				***1,68 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***3,57 meses/hombre***
		2,00	*2,40*	*0,60*	*0,70*	
		$PEM_E = [(1{,}80 \cdot 2{,}00) + (0{,}90 \cdot 2{,}40) + (1{,}40 \cdot 0{,}60) - 1{,}50] \cdot 0{,}70$				

Tabla D.4. Resultados de aplicar los métodos del modelo al proyecto P4.

#	Método	Aplicación del Método				Esfuerzo Estimado
P5	*Fórmula Lineal*	$PEM_L = (0{,}80 \cdot 3) + (1{,}10 \cdot 2) - (1{,}20 \cdot 2) - (0{,}30 \cdot 5) - (0{,}70 \cdot 2) + (1{,}80 \cdot 3) - (0{,}90 \cdot 1) + (1{,}86 \cdot 5) - 3{,}30$				***9,80 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***7,58 meses/hombre***
		3,70	2,40	3,80	0,60	
		$PEM_E = [(1{,}80 \cdot 3{,}70) + (0{,}90 \cdot 2{,}40) + (1{,}40 \cdot 3{,}80) - 1{,}50] \cdot 0{,}60$				

Tabla D.5. Resultados de aplicar los métodos del modelo al proyecto P5.

#	Método	Aplicación del Método				Esfuerzo Estimado
P6	*Fórmula Lineal*	$PEM_L = (0{,}80 \cdot 4) + (1{,}10 \cdot 1) - (1{,}20 \cdot 1) - (0{,}30 \cdot 2) - (0{,}70 \cdot 1) + (1{,}80 \cdot 1) - (0{,}90 \cdot 5) + (1{,}86 \cdot 5) - 3{,}30$				***5,10 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***6,07 meses/hombre***
		1,00	*0,50*	*3,80*	*1,00*	
		$PEM_E = [(1{,}80 \cdot 1{,}00) + (0{,}90 \cdot 0{,}50) + (1{,}40 \cdot 3{,}80) - 1{,}50] \cdot 1{,}00$				

Tabla D.6. Resultados de aplicar los métodos del modelo al proyecto P6.

#	Método	Aplicación del Método				Esfuerzo Estimado
P7	*Fórmula Lineal*	$PEM_L = (0{,}80 \cdot 3) + (1{,}10 \cdot 2) - (1{,}20 \cdot 1) - (0{,}30 \cdot 4) - (0{,}70 \cdot 1) + (1{,}80 \cdot 1) - (0{,}90 \cdot 2) + (1{,}86 \cdot 3) - 3{,}30$				***3,78 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***5,91 meses/hombre***
		3,70	*2,40*	*0,80*	*0,70*	
		$PEM_E = [(1{,}80 \cdot 3{,}70) + (0{,}90 \cdot 2{,}40) + (1{,}40 \cdot 0{,}80) - 1{,}50] \cdot 0{,}70$				

Tabla D.7. Resultados de aplicar los métodos del modelo al proyecto P7.

#	Método	Aplicación del Método				Esfuerzo Estimado
P8	*Fórmula Lineal*	$PEM_L = (0{,}80 \cdot 1) + (1{,}10 \cdot 4) - (1{,}20 \cdot 1) - (0{,}30 \cdot 3) - (0{,}70 \cdot 2) + (1{,}80 \cdot 1) - (0{,}90 \cdot 1) + (1{,}86 \cdot 3) - 3{,}30$				***4,88 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***3,06 meses/hombre***
		2,00	*2,40*	*0,60*	*0,60*	
		$PEM_E = [(1{,}80 \cdot 2{,}00) + (0{,}90 \cdot 2{,}40) + (1{,}40 \cdot 0{,}60) - 1{,}50] \cdot 0{,}60$				

Tabla D.8. Resultados de aplicar los métodos del modelo al proyecto P8.

#	Método	Aplicación del Método				Esfuerzo Estimado
P9	*Fórmula Lineal*	$PEM_L = (0{,}80 \cdot 5) + (1{,}10 \cdot 1) - (1{,}20 \cdot 1) - (0{,}30 \cdot 3) - (0{,}70 \cdot 3) + (1{,}80 \cdot 3) - (0{,}90 \cdot 4) + (1{,}86 \cdot 5) - 3{,}30$				***8,70 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***7,66 meses/hombre***
		2,00	2,40	3,80	0,80	
		$PEM_E = [(1{,}80 \cdot 2{,}00) + (0{,}90 \cdot 2{,}40) + (1{,}40 \cdot 3{,}80) - 1{,}50] \cdot 0{,}80$				

Tabla D.9. Resultados de aplicar los métodos del modelo al proyecto P9.

#	Método	Aplicación del Método				Esfuerzo Estimado
P10	*Fórmula Lineal*	$PEM_L = (0{,}80 \cdot 4) + (1{,}10 \cdot 1) - (1{,}20 \cdot 2) - (0{,}30 \cdot 2) - (0{,}70 \cdot 1) + (1{,}80 \cdot 1) - (0{,}90 \cdot 4) + (1{,}86 \cdot 3) - 3{,}30$				***1,08 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***1,50 meses/hombre***
		1,00	*0,50*	*0,80*	*0,80*	
		$PEM_E = [(1{,}80 \cdot 1{,}00) + (0{,}90 \cdot 0{,}50) + (1{,}40 \cdot 0{,}80) - 1{,}50] \cdot 0{,}80$				

Tabla D.10. Resultados de aplicar los métodos del modelo al proyecto P10.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P11</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0{,}80 \cdot 3) + (1{,}10 \cdot 3) - (1{,}20 \cdot 1) - (0{,}30 \cdot 5) - (0{,}70 \cdot 3) + (1{,}80 \cdot 4) - (0{,}90 \cdot 5) + (1{,}86 \cdot 5) - 3{,}30$</td><td>9,60 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">12,64 meses/hombre</td></tr>
<tr><td>3,70</td><td>2,40</td><td>3,80</td><td>1,00</td></tr>
<tr><td colspan="4">$PEM_E = [(1{,}80 \cdot 3{,}70) + (0{,}90 \cdot 2{,}40) + (1{,}40 \cdot 3{,}80) - 1{,}50] \cdot 1{,}00$</td></tr>
</table>

Tabla D.11. Resultados de aplicar los métodos del modelo al proyecto P11.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P12</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0{,}80 \cdot 1) + (1{,}10 \cdot 2) - (1{,}20 \cdot 1) - (0{,}30 \cdot 5) - (0{,}70 \cdot 2) + (1{,}80 \cdot 2) - (0{,}90 \cdot 3) + (1{,}86 \cdot 5) - 3{,}30$</td><td>5,80 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">5,63 meses/hombre</td></tr>
<tr><td>2,00</td><td>2,40</td><td>2,70</td><td>0,70</td></tr>
<tr><td colspan="4">$PEM_E = [(1{,}80 \cdot 2{,}00) + (0{,}90 \cdot 2{,}40) + (1{,}40 \cdot 2{,}70) - 1{,}50] \cdot 0{,}70$</td></tr>
</table>

Tabla D.12. Resultados de aplicar los métodos del modelo al proyecto P12.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P13</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0{,}80 \cdot 1) + (1{,}10 \cdot 1) - (1{,}20 \cdot 2) - (0{,}30 \cdot 1) - (0{,}70 \cdot 2) + (1{,}80 \cdot 5) - (0{,}90 \cdot 5) + (1{,}86 \cdot 3) - 3{,}30$</td><td>4,58 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">3,17 meses/hombre</td></tr>
<tr><td>2,00</td><td>0,25</td><td>0,60</td><td>1,00</td></tr>
<tr><td colspan="4">$PEM_E = [(1{,}80 \cdot 2{,}00) + (0{,}90 \cdot 0{,}25) + (1{,}40 \cdot 0{,}60) - 1{,}50] \cdot 1{,}00$</td></tr>
</table>

Tabla D.13. Resultados de aplicar los métodos del modelo al proyecto P13.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P14</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0{,}80 \cdot 3) + (1{,}10 \cdot 3) - (1{,}20 \cdot 1) - (0{,}30 \cdot 3) - (0{,}70 \cdot 3) + (1{,}80 \cdot 4) - (0{,}90 \cdot 2) + (1{,}86 \cdot 3) - 3{,}30$</td><td>9,18 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">5,91 meses/hombre</td></tr>
<tr><td>3,70</td><td>2,40</td><td>0,80</td><td>0,70</td></tr>
<tr><td colspan="4">$PEM_E = [(1{,}80 \cdot 3{,}70) + (0{,}90 \cdot 2{,}40) + (1{,}40 \cdot 0{,}80) - 1{,}50] \cdot 0{,}70$</td></tr>
</table>

Tabla D.14. Resultados de aplicar los métodos del modelo al proyecto P14.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P15</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0{,}80 \,.\, 1) + (1{,}10 \,.\, 1) - (1{,}20 \,.\, 1) - (0{,}30 \,.\, 2) - (0{,}70 \,.\, 1) + (1{,}80 \,.\, 2) - (0{,}90 \,.\, 2) + (1{,}86 \,.\, 3) - 3{,}30$</td><td>3,48 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">1,11 meses/hombre</td></tr>
<tr><td>1,00</td><td>0,50</td><td>0,60</td><td>0,70</td></tr>
<tr><td colspan="4">$PEM_E = [\,(1{,}80 \,.\, 1{,}00) + (0{,}90 \,.\, 0{,}50) + (1{,}40 \,.\, 0{,}60) - 1{,}50\,] \,.\, 0{,}70$</td></tr>
</table>

Tabla D.15. Resultados de aplicar los métodos del modelo al proyecto P15.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P16</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0{,}80 \,.\, 3) + (1{,}10 \,.\, 4) - (1{,}20 \,.\, 1) - (0{,}30 \,.\, 3) - (0{,}70 \,.\, 2) + (1{,}80 \,.\, 2) - (0{,}90 \,.\, 1) + (1{,}86 \,.\, 5) - 3{,}30$</td><td>12,00 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">7,58 meses/hombre</td></tr>
<tr><td>3,70</td><td>2,40</td><td>3,80</td><td>0,60</td></tr>
<tr><td colspan="4">$PEM_E = [\,(1{,}80 \,.\, 3{,}70) + (0{,}90 \,.\, 2{,}40) + (1{,}40 \,.\, 3{,}80) - 1{,}50\,] \,.\, 0{,}60$</td></tr>
</table>

Tabla D.16. Resultados de aplicar los métodos del modelo al proyecto P16.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P17</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0{,}80 \,.\, 3) + (1{,}10 \,.\, 3) - (1{,}20 \,.\, 1) - (0{,}30 \,.\, 5) - (0{,}70 \,.\, 3) + (1{,}80 \,.\, 2) - (0{,}90 \,.\, 5) + (1{,}86 \,.\, 3) - 3{,}30$</td><td>2,28 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">8,44 meses/hombre</td></tr>
<tr><td>3,70</td><td>2,40</td><td>0,80</td><td>1,00</td></tr>
<tr><td colspan="4">$PEM_E = [\,(1{,}80 \,.\, 3{,}70) + (0{,}90 \,.\, 2{,}40) + (1{,}40 \,.\, 0{,}80) - 1{,}50\,] \,.\, 1{,}00$</td></tr>
</table>

Tabla D.17. Resultados de aplicar los métodos del modelo al proyecto P17.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P18</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0{,}80 \,.\, 1) + (1{,}10 \,.\, 2) - (1{,}20 \,.\, 1) - (0{,}30 \,.\, 3) - (0{,}70 \,.\, 2) + (1{,}80 \,.\, 2) - (0{,}90 \,.\, 2) + (1{,}86 \,.\, 3) - 3{,}30$</td><td>3,58 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">3,57 meses/hombre</td></tr>
<tr><td>2,00</td><td>2,40</td><td>0,60</td><td>0,70</td></tr>
<tr><td colspan="4">$PEM_E = [\,(1{,}80 \,.\, 2{,}00) + (0{,}90 \,.\, 2{,}40) + (1{,}40 \,.\, 0{,}60) - 1{,}50\,] \,.\, 0{,}70$</td></tr>
</table>

Tabla D.18. Resultados de aplicar los métodos del modelo al proyecto P18.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P19</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0,80 . 3) + (1,10 . 3) - (1,20 . 2) - (0,30 . 3) - (0,70 . 3) + (1,80 . 2) - (0,90 . 4) + (1,86 . 5) - 3,30$</td><td>6,30 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">10,11 meses/hombre</td></tr>
<tr><td>3,70</td><td>2,40</td><td>3,80</td><td>0,80</td></tr>
<tr><td colspan="4">$PEM_E = [(1,80 . 3,70) + (0,90 . 2,40) + (1,40 . 3,80) - 1,50] . 0,80$</td></tr>
</table>

Tabla D.19. Resultados de aplicar los métodos del modelo al proyecto P19.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P20</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0,80 . 5) + (1,10 . 4) - (1,20 . 2) - (0,30 . 5) - (0,70 . 3) + (1,80 . 5) - (0,90 . 5) + (1,86 . 3) - 3,30$</td><td>9,18 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">8,44 meses/hombre</td></tr>
<tr><td>3,70</td><td>2,40</td><td>0,80</td><td>1,00</td></tr>
<tr><td colspan="4">$PEM_E = [(1,80 . 3,70) + (0,90 . 2,40) + (1,40 . 0,80) - 1,50] . 1,00$</td></tr>
</table>

Tabla D.20. Resultados de aplicar los métodos del modelo al proyecto P20.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P21</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0,80 . 1) + (1,10 . 3) - (1,20 . 1) - (0,30 . 1) - (0,70 . 1) + (1,80 . 4) - (0,90 . 4) + (1,86 . 5) - 3,30$</td><td>11,50 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">7,33 meses/hombre</td></tr>
<tr><td>3,70</td><td>0,25</td><td>2,70</td><td>0,80</td></tr>
<tr><td colspan="4">$PEM_E = [(1,80 . 3,70) + (0,90 . 0,25) + (1,40 . 2,70) - 1,50] . 0,80$</td></tr>
</table>

Tabla D.21. Resultados de aplicar los métodos del modelo al proyecto P21.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P22</td><td>Fórmula Lineal</td><td colspan="4">$PEM_L = (0,80 . 3) + (1,10 . 1) - (1,20 . 2) - (0,30 . 5) - (0,70 . 4) + (1,80 . 3) - (0,90 . 2) + (1,86 . 5) - 3,30$</td><td>6,40 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">6,71 meses/hombre</td></tr>
<tr><td>2,00</td><td>2,40</td><td>3,80</td><td>0,70</td></tr>
<tr><td colspan="4">$PEM_E = [(1,80 . 2,00) + (0,90 . 2,40) + (1,40 . 3,80) - 1,50] . 0,70$</td></tr>
</table>

Tabla D.22. Resultados de aplicar los métodos del modelo al proyecto P22.

#	Método	Aplicación del Método				Esfuerzo Estimado
P23	*Fórmula Lineal*	$PEM_L = (0{,}80 \,.\, 4) + (1{,}10 \,.\, 2) - (1{,}20 \,.\, 1) - (0{,}30 \,.\, 3) - (0{,}70 \,.\, 2) + (1{,}80 \,.\, 3) - (0{,}90 \,.\, 4) + (1{,}86 \,.\, 5) - 3{,}30$)				***9,70 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***10,11 meses/hombre***
		3,70	*2,40*	*3,80*	*0,80*	
		$PEM_E = [\,(1{,}80 \,.\, 3{,}70) + (0{,}90 \,.\, 2{,}40) + (1{,}40 \,.\, 3{,}80) - 1{,}50\,] \,.\, 0{,}80$				

Tabla D.23. Resultados de aplicar los métodos del modelo al proyecto P23.

#	Método	Aplicación del Método				Esfuerzo Estimado
P24	*Fórmula Lineal*	$PEM_L = (0{,}80 \,.\, 4) + (1{,}10 \,.\, 3) - (1{,}20 \,.\, 1) - (0{,}30 \,.\, 2) - (0{,}70 \,.\, 1) + (1{,}80 \,.\, 4) - (0{,}90 \,.\, 5) + (1{,}86 \,.\, 5) - 3{,}30$)				***12,70 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***10,93 meses/hombre***
		3,70	*0,50*	*3,80*	*1,00*	
		$PEM_E = [\,(1{,}80 \,.\, 3{,}70) + (0{,}90 \,.\, 0{,}50) + (1{,}40 \,.\, 3{,}80) - 1{,}50\,] \,.\, 1{,}00$				

Tabla D.24. Resultados de aplicar los métodos del modelo al proyecto P24.

#	Método	Aplicación del Método				Esfuerzo Estimado
P25	*Fórmula Lineal*	$PEM_L = (0{,}80 \,.\, 3) + (1{,}10 \,.\, 3) - (1{,}20 \,.\, 2) - (0{,}30 \,.\, 1) - (0{,}70 \,.\, 2) + (1{,}80 \,.\, 5) - (0{,}90 \,.\, 5) + (1{,}86 \,.\, 3) - 3{,}30$)				***8,38 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***6,51 meses/hombre***
		3,70	*0,25*	*0,80*	*1,00*	
		$PEM_E = [\,(1{,}80 \,.\, 3{,}70) + (0{,}90 \,.\, 0{,}25) + (1{,}40 \,.\, 0{,}80) - 1{,}50\,] \,.\, 1{,}00$				

Tabla D.25. Resultados de aplicar los métodos del modelo al proyecto P25.

#	Método	Aplicación del Método				Esfuerzo Estimado
P26	*Fórmula Lineal*	$PEM_L = (0{,}80 \,.\, 1) + (1{,}10 \,.\, 2) - (1{,}20 \,.\, 1) - (0{,}30 \,.\, 5) - (0{,}70 \,.\, 3) + (1{,}80 \,.\, 2) - (0{,}90 \,.\, 3) + (1{,}86 \,.\, 5) - 3{,}30$)				***5,10 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***5,63 meses/hombre***
		2,00	*2,40*	*2,70*	*0,70*	
		$PEM_E = [\,(1{,}80 \,.\, 2{,}00) + (0{,}90 \,.\, 2{,}40) + (1{,}40 \,.\, 2{,}70) - 1{,}50\,] \,.\, 0{,}70$				

Tabla D.26. Resultados de aplicar los métodos del modelo al proyecto P26.

#	Método	Aplicación del Método				Esfuerzo Estimado
P27	*Fórmula Lineal*	$PEM_L = (0{,}80 \,.\, 1) + (1{,}10 \,.\, 1) - (1{,}20 \,.\, 2) - (0{,}30 \,.\, 4) - (0{,}70 \,.\, 3) + (1{,}80 \,.\, 4) - (0{,}90 \,.\, 3) + (1{,}86 \,.\, 5) - 3{,}30)$				***6,70 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***5,63 meses/hombre***
		2,00	*2,40*	*2,70*	*0,70*	
		$PEM_E = [(1{,}80 \,.\, 2{,}00) + (0{,}90 \,.\, 2{,}40) + (1{,}40 \,.\, 2{,}70) - 1{,}50] \,.\, 0{,}70$				

Tabla D.27. Resultados de aplicar los métodos del modelo al proyecto P27.

#	Método	Aplicación del Método				Esfuerzo Estimado
P28	*Fórmula Lineal*	$PEM_L = (0{,}80 \,.\, 1) + (1{,}10 \,.\, 3) - (1{,}20 \,.\, 1) - (0{,}30 \,.\, 2) - (0{,}70 \,.\, 2) + (1{,}80 \,.\, 4) - (0{,}90 \,.\, 5) + (1{,}86 \,.\, 5) - 3{,}30)$				***9,60 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***9,39 meses/hombre***
		3,70	*0,50*	*2,70*	*1,00*	
		$PEM_E = [(1{,}80 \,.\, 3{,}70) + (0{,}90 \,.\, 0{,}50) + (1{,}40 \,.\, 2{,}70) - 1{,}50] \,.\, 1{,}00$				

Tabla D.28. Resultados de aplicar los métodos del modelo al proyecto P28.

#	Método	Aplicación del Método				Esfuerzo Estimado
P29	*Fórmula Lineal*	$PEM_L = (0{,}80 \,.\, 3) + (1{,}10 \,.\, 2) - (1{,}20 \,.\, 1) - (0{,}30 \,.\, 3) - (0{,}70 \,.\, 3) + (1{,}80 \,.\, 5) - (0{,}90 \,.\, 3) + (1{,}86 \,.\, 2) - 3{,}30)$				***7,12 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***5,91 meses/hombre***
		3,70	*2,40*	*0,80*	*0,70*	
		$PEM_E = [(1{,}80 \,.\, 3{,}70) + (0{,}90 \,.\, 2{,}40) + (1{,}40 \,.\, 0{,}80) - 1{,}50] \,.\, 0{,}70$				

Tabla D.29. Resultados de aplicar los métodos del modelo al proyecto P29.

#	Método	Aplicación del Método				Esfuerzo Estimado
P30	*Fórmula Lineal*	$PEM_L = (0{,}80 \,.\, 5) + (1{,}10 \,.\, 3) - (1{,}20 \,.\, 2) - (0{,}30 \,.\, 5) - (0{,}70 \,.\, 4) + (1{,}80 \,.\, 4) - (0{,}90 \,.\, 4) + (1{,}86 \,.\, 5) - 3{,}30)$				***10,20 meses/hombre***
	Método Empírico	CNEG	CDAT	CMOD	AEXP	***10,11 meses/hombre***
		3,70	*2,40*	*3,80*	*0,80*	
		$PEM_E = [(1{,}80 \,.\, 3{,}70) + (0{,}90 \,.\, 2{,}40) + (1{,}40 \,.\, 3{,}80) - 1{,}50] \,.\, 0{,}80$				

Tabla D.30. Resultados de aplicar los métodos del modelo al proyecto P30.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P31</td><td>Fórmula Lineal</td><td colspan="4">PEM_L = (0,80 . 1) + (1,10 . 1) - (1,20 . 1) - (0,30 . 2) - (0,70 . 1) + (1,80 . 1) - (0,90 . 2) + (1,86 . 3) - 3,30)</td><td>1,68 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">1,11 meses/hombre</td></tr>
<tr><td>1,00</td><td>0,50</td><td>0,60</td><td>0,70</td></tr>
<tr><td colspan="4">PEM_E = [(1,80 . 1,00) + (0,90 . 0,50) + (1,40 . 0,60) - 1,50] . 0,70</td></tr>
</table>

Tabla D.31. Resultados de aplicar los métodos del modelo al proyecto P31.

<table>
<tr><th>#</th><th>Método</th><th colspan="4">Aplicación del Método</th><th>Esfuerzo Estimado</th></tr>
<tr><td rowspan="4">P32</td><td>Fórmula Lineal</td><td colspan="4">PEM_L = (0,80 . 3) + (1,10 . 2) - (1,20 . 1) - (0,30 . 2) - (0,70 . 1) + (1,80 . 1) - (0,90 . 1) + (1,86 . 2) - 3,30)</td><td>3,42 meses/hombre</td></tr>
<tr><td rowspan="3">Método Empírico</td><td>CNEG</td><td>CDAT</td><td>CMOD</td><td>AEXP</td><td rowspan="3">4,04 meses/hombre</td></tr>
<tr><td>3,70</td><td>0,50</td><td>0,80</td><td>0,60</td></tr>
<tr><td colspan="4">PEM_E = [(1,80 . 3,70) + (0,90 . 0,50) + (1,40 . 0,80) - 1,50] . 0,60</td></tr>
</table>

Tabla D.32. Resultados de aplicar los métodos del modelo al proyecto P32.

ANEXO E: APLICACIÓN DEL MODELO DMCoMo EN PROYECTOS DE EXPLOTACIÓN PARA SU VALIDACIÓN

En este anexo se indican los resultados de aplicar el modelo para la estimación de esfuerzo DMCoMo propuesto en [Marbán *et al.*, 2008].

El modelo DMCoMo se aplica sobre los primeros treinta y dos proyectos de explotación de información presentados en el Anexo B, los cuales también fueron utilizados para la validación de os dos métodos del modelo de estimación propuesto en este trabajo de tesis.

Primero se lleva a cabo una la caracterización de los proyectos teniendo en cuenta los factores de costo del modelo DMCoMo; dicha caracterizan se indica en las tablas E.1, E.2 y E.3. Posteriormente se aplican las dos fórmulas del modelo (MM8 y MM23) obteniendo los resultados para cada proyecto se muestran en las Tablas E.4 y E.5. En dichas tablas también se incluye la comparación con el esfuerzo real del proyecto y el cálculo del error correspondiente.

Finalmente, para ilustrar las diferencias entre los resultados de DMCoMo, el esfuerzo real de los proyectos y cada uno de los métodos de estimación del modelo propuesto se muestra el gráfico Boxplot de la Figura E.1.

#	NTAB	NTUP	NATR	DISP	PNUL	DMOD	DEXT	NMOD	TMOD	MTUP	MATR	MTEC	NFUN	SCOM	TOOL	COMP	NFOR	NDEP	DOCU	SITE	KDAT	ADIR	MFAM
P1	1	1	7	1	0	1	1	1	0	1	1	1	3	1	1	3	3	4	5	3	4	1	3
P2	0	1	1	1	1	4	0	2	1	1	2	1	1	1	1	5	3	2	2	1	3	1	5
P3	0	1	1	1	1	1	0	2	1	1	2	1	1	1	1	0	3	2	3	1	3	1	5
P4	3	5	5	2	2	2	1	3	3	3	3	5	3	0	1	2	3	1	2	0	2	4	3
P5	1	3	3	2	1	5	2	3	1	3	3	3	2	1	1	1	1	4	2	2	4	2	1
P6	0	1	1	1	2	1	2	2	1	1	2	1	1	1	1	0	1	1	2	1	1	1	5
P7	1	1	1	1	2	1	2	2	3	1	2	4	1	1	1	1	3	3	2	3	2	2	3
P8	2	0	3	4	1	0	0	1	1	0	3	1	0	1	1	4	2	0	2	0	1	6	0
P9	0	1	1	1	1	1	2	2	4	1	2	1	1	1	1	3	3	4	5	3	3	1	4

Tabla E.1. Caracterización de los proyectos para el modelo de estimación DMCoMo (P1 a P9).

#	NTAB	NTUP	NATR	DISP	PNUL	DMOD	DEXT	NMOD	TMOD	MTUP	MATR	MTEC	NFUN	SCOM	TOOL	COMP	NFOR	NDEP	DOCU	SITE	KDAT	ADIR	MFAM
P10	0	1	1	1	1	0	2	2	4	1	2	4	2	1	1	2	1	2	5	3	2	1	4
P11	0	1	1	2	1	5	2	2	3	1	1	3	1	2	1	1	2	3	3	2	2	3	4
P12	0	1	1	2	2	3	2	2	2	1	1	3	1	2	1	1	2	3	3	2	2	3	2
P14	0	1	1	1	1	5	2	2	2	1	1	3	1	2	1	1	2	3	3	2	2	3	1
P15	0	1	1	1	1	4	2	2	3	1	1	3	1	2	1	1	2	4	3	2	2	2	1
P16	1	1	0	2	4	6	2	1	0	1	1	1	2	3	6	0	3	0	1	0	3	3	3
P17	1	1	1	1	1	2	2	2	1	1	5	2	1	1	1	3	2	2	2	1	3	2	3
P18	0	1	1	1	1	1	2	2	1	1	3	4	1	1	2	1	3	2	3	1	1	2	5
P19	3	0	4	2	2	1	1	2	2	0	4	2	1	1	4	4	4	2	3	0	3	5	2
P20	3	3	12	2	0	0	50	0	1	5	5	2	1	1	0	0	40	1	20	1	40	1	4
P21	1	1	5	1	1	5	2	2	4	1	1	2	1	1	1	5	2	2	2	1	1	1	1
P22	3	5	10	3	0	1	1	1	1	4	4	5	3	0	1	8	1	3	1	1	1	3	5
P14	0	1	1	1	1	5	2	2	2	1	1	3	1	2	1	1	2	3	3	2	2	3	1
P15	0	1	1	1	1	4	2	2	3	1	1	3	1	2	1	1	2	4	3	2	2	2	1
P16	1	1	0	2	4	6	2	1	0	1	1	1	2	3	6	0	3	0	1	0	3	3	3
P17	1	1	1	1	1	2	2	2	1	1	5	2	1	1	1	3	2	2	2	1	3	2	3
P18	0	1	1	1	1	1	2	2	1	1	3	4	1	1	2	1	3	2	3	1	1	2	5
P19	3	0	4	2	2	1	1	2	2	0	4	2	1	1	4	4	4	2	3	0	3	5	2
P20	3	3	12	2	0	0	50	0	1	5	5	2	1	1	0	0	40	1	20	1	40	1	4
P21	1	1	5	1	1	5	2	2	4	1	1	2	1	1	1	5	2	2	2	1	1	1	1
P22	3	5	10	3	0	1	1	1	1	4	4	5	3	0	1	8	1	3	1	1	1	3	5
P23	1	1	1	1	1	2	1	1	4	1	3	3	1	2	1	5	1	2	3	1	1	2	5

Tabla E.2. Caracterización de los proyectos para el modelo de estimación DMCoMo (P10 a P23).

#	NTAB	NTUP	NATR	DISP	PNUL	DMOD	DEXT	NMOD	TMOD	MTUP	MATR	MTEC	NFUN	SCOM	TOOL	COMP	NFOR	NDEP	DOCU	SITE	KDAT	ADIR	MFAM
P24	1	1	1	3	2	1	1	1	4	1	2	3	1	1	1	5	2	2	3	0	3	5	2
P25	1	1	1	2	1	2	2	1	2	1	4	2	2	3	1	0	1	1	2	1	4	1	4
P26	1	2	1	2	2	2	1	1	1	2	3	2	1	2	0	2	1	2	2	1	1	1	1
P27	1	1	1	1	3	1	1	1	1	1	2	2	1	2	1	5	2	3	1	1	1	3	5
P28	1	1	1	1	2	1	1	1	1	1	1	2	1	2	0	0	1	2	3	1	1	2	5
P29	1	1	1	2	1	1	1	1	2	1	2	4	1	2	1	1	1	2	3	0	3	5	2
P30	1	2	1	2	3	2	1	2	1	2	3	2	3	3	1	1	2	1	2	1	4	1	4
P31	1	1	1	2	1	1	1	1	1	1	4	2	1	1	1	1	1	2	2	1	1	1	1
P32	1	1	1	1	1	1	1	1	2	1	4	2	1	1	1	1	1	3	1	1	1	3	5

Tabla E.3. Caracterización de los proyectos para el modelo de estimación DMCoMo (P24 a P32).

#	Esfuerzo Real ER	Fórmula MM8			Fórmula MM23		
		Esfuerzo calculado MM8	Error ER – MM8	Error Relativo $\frac{ER - MM8}{ER}$	Esfuerzo calculado MM23	Error ER – MM23	Error Relativo $\frac{ER - MM23}{ER}$
P1	*2,41*	*84,23*	- 81,82	- 3.395%	*94,88*	- 92,47	- 3837%
P2	*7,00*	*67,16*	- 60,16	- 859%	*51,84*	- 44,84	- 641%
P3	*1,64*	*67,16*	- 65,52	- 3.995%	*68,07*	- 66,43	- 4.051%
P4	*3,65*	*118,99*	- 115,34	- 3.160%	*111,47*	- 107,82	- 2.954%
P5	*9,35*	*110,92*	- 101,57	- 1.086%	*122,52*	- 113,17	- 1.210%
P6	*11,63*	*80,27*	- 68,64	- 590%	*81,36*	- 69,73	- 600%
P7	*6,73*	*96,02*	- 89,29	- 1.327%	*92,49*	- 85,76	- 1.274%
P8	*5,40*	*116,87*	- 111,47	- 2.064%	*89,68*	- 84,28	- 1.561%
P9	*8,38*	*97,63*	- 89,25	- 1.065%	*98,74*	- 90,36	- 1.078%
P10	*1,56*	*105,32*	- 103,76	- 6.651%	*103,13*	- 101,57	- 6.511%
P11	*9,70*	*94,39*	- 84,69	- 873%	*77,03*	- 67,33	- 694%
P12	*5,24*	*93,69*	- 88,45	- 1.688%	*85,74*	- 80,50	- 1.536%
P13	*5,00*	*91,05*	- 86,05	- 1.721%	*93,08*	- 88,08	- 1.762%
P14	*8,97*	*92,17*	- 83,20	- 928%	*78,20*	- 69,23	- 772%
P15	*2,81*	*99,43*	- 96,62	- 3.438%	*93,57*	- 90,76	- 3.230%

Tabla E.4. Comparación del Esfuerzo Real con los calculados por el método DMCoMo en meses/hombre (de P1 a P15).

#	Esfuerzo Real ER	Fórmula MM8			Fórmula MM23		
		Esfuerzo calculado MM8	Error ER – MM8	Error Relativo $\frac{ER - MM8}{ER}$	Esfuerzo calculado MM23	Error ER – MM23	Error Relativo $\frac{ER - MM23}{ER}$
P16	*11,80*	*71,54*	- 59,74	- 506%	*35,59*	- 23,79	- 202%
P17	*2,79*	*99,19*	- 96,40	- 3.455%	*91,12*	- 88,33	- 3.166%
P18	*3,88*	*77,20*	- 73,32	- 1.890%	*60,66*	- 56,78	- 1.464%
P19	*5,70*	*112,89*	- 107,19	- 1.881%	*69,90*	- 64,20	- 1.126%
P20	*8,54*	*121,41*	- 112,87	- 1.322%	*81,81*	- 73,27	- 858%
P21	*10,61*	*120,59*	- 109,98	- 1.037%	*99,45*	- 88,84	- 837%
P22	*6,88*	*129,44*	- 122,56	- 1.781%	*130,73*	- 123,85	- 1.800%
P23	*11,20*	*106,31*	- 95,11	- 849%	*86,93*	- 75,73	- 676%
P24	*9,70*	*117,26*	- 107,56	- 1.109%	*92,03*	- 82,33	- 849%
P25	*7,30*	*107,19*	- 99,89	- 1.368%	*111,05*	- 103,75	- 1.421%
P26	*5,31*	*102,43*	- 97,12	- 1.829%	*117,39*	- 112,08	- 2.111%
P27	*6,10*	*76,08*	- 69,98	- 1.147%	*66,08*	- 59,98	- 983%
P28	*10,00*	*75,31*	- 65,31	- 653%	*78,27*	- 68,27	- 683%
P29	*6,43*	*101,80*	- 95,37	- 1.483%	*83,52*	- 77,09	- 1.199%
P30	*9,80*	*88,77*	- 78,97	- 806%	*101,39*	- 91,59	- 935%
P31	*1,50*	*107,05*	- 105,55	- 7.036%	*114,72*	- 113,22	- 7.548%
P32	*3,78*	*96,41*	- 92,63	- 2.451%	*92,47*	- 88,69	- 2.346%

Tabla E.5. Comparación del Esfuerzo Real con los calculados por el método DMCoMo en meses/hombre, (de P16 a P32).

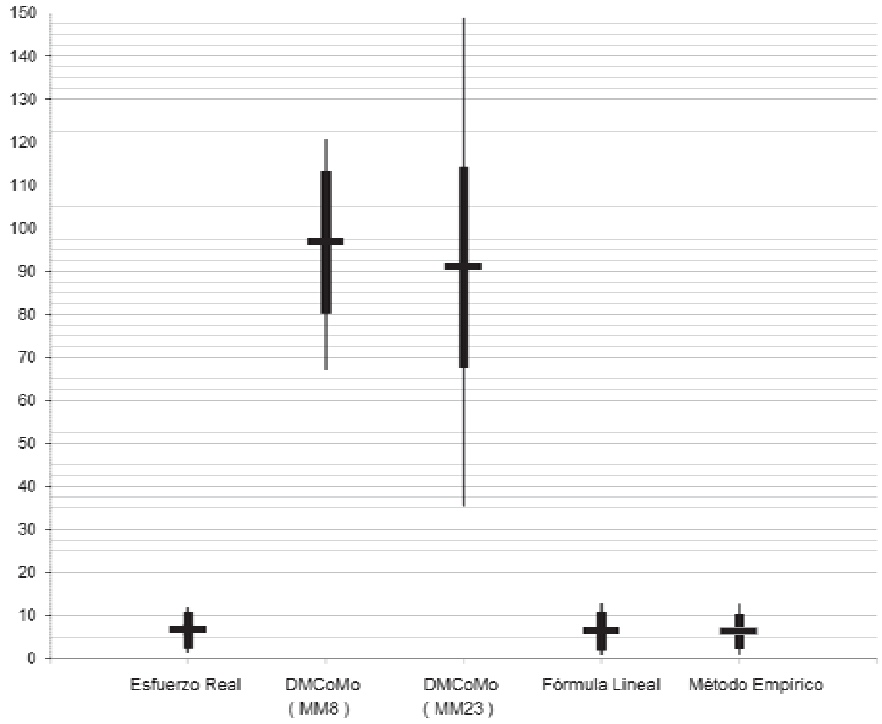


Figura E.1. Gráfico boxplot comprando las estimaciones de DMCoMo con el esfuerzo real y los estimados por del modelo propuesto.

ANEXO F: DESCRIPCIÓN DE LA PRUEBA DE RANGOS CON SIGNO DE WILCOXON

En este anexo se describe el funcionamiento de la prueba de rangos con signo de Wilcoxon [1945] la cual es aplicada en el capítulo 5 correspondiente a la validación de los modelos propuestos en este trabajo de tesis.

- **Objetivo de la Prueba:**

  La prueba de rangos con signo de Wilcoxon permite analizar la similitud entre un conjunto de datos muestrales apareados donde cada elemento de la población posee un valor experimental que se desea comprobar y un valor de referencia o de control [Guardía-Olmos *et al.,* 2007]. Como resultado de la prueba, es posible aceptar o refutar una hipótesis nula (H_0) la cual indica que la población de diferencias entre los valores experimentales y los valores de control tienen una mediana de cero [Triola, 2013]. En otros términos, al comprobar la hipótesis nula mediante esta prueba no paramétrica, se corrobora que no existen diferencias significativas entre los datos experimentales y los datos de control.

  Por ejemplo, esta prueba puede ser utilizada para medir la efectividad de un medicamento para dormir, a los efectos de comparar las horas de sueño de los pacientes cuando se les suministró el medicamento (datos experimentales) y las horas de sueño de los mismos pacientes cuando se les suministró un placebo (datos de control). Para este ejemplo, si la hipótesis nula es aceptada, entonces el medicamento no produce ninguna diferencia en el sueño de los pacientes; en cambio, si es rechazada, se puede comprobar que produce alguna diferencia (pudiendo resultar ésta una mejoría o no).

- **Requisitos de la Prueba:**

  Para poder llevar a cabo esta prueba se debe cumplir con los siguientes requisitos:

  i. Todos los valores utilizados deben ser continuos.
  ii. Para cada elemento se encuentra disponible dos valores (datos apareados), los cuales han sido seleccionados en forma aleatoria.
  iii. La población de las diferencias (calculadas a partir de los pares de datos) tiene una distribución que es aproximadamente simétrica, lo que significa que la mitad izquierda de su histograma es, de manera aproximada, una imagen de espejo de la mitad derecha. Por consiguiente, los datos no necesitan tener una distribución normal.

- **Nivel de Significancia y Grado de Confianza de la Prueba:**

  Al llevar a cabo esta prueba es posible determinar el máximo grado de error (o de riesgo de cometer un error). En este sentido, antes de aplicar la prueba se debe especificar el valor del "nivel de significancia" (también denominado "nivel de significación") seleccionado, que se suele indica con la letra griega "α". Cabe recordar que el nivel de significancia es un concepto estadístico que define la probabilidad de rechazar la hipótesis nula cuando esta es cierta [Iglesias Pérez, S/A]. Por consiguiente, este nivel se encuentra asociado a la probabilidad de cometer un error en la prueba y suele ser un valor pequeño (siendo 0,05 el más utilizado).

  Otra forma de representar este concepto es mediante el "grado de confianza"; el cual define la probabilidad de no cometer un error al rechazar la hipótesis nula, por lo que dicho grado es calculado como el opuesto del nivel de significancia ($GradoConfianza = 1 - \alpha$). Por ejemplo, si se utiliza un nivel de significancia igual a 0,05; entonces el grado de confianza asociado será del 95%.

- **Procedimiento de la Prueba:**

  A partir de los datos recolectados (población de elementos con los valores experimentales y de control) y una vez seleccionado el nivel de significancia (α), se deben llevar a cabo los siguientes pasos esta prueba [Triola, 2013]:

  1) Para cada elemento de la población, calcular la diferencia restando el valor control al valor de experimental ($Diferencia = ValorExperimental - ValorControl$). En este paso se debe mantener los signos, pero descartar cualquier elemento donde la diferencia sea igual a cero.

  2) Utilizando las diferencias obtenidas en el paso anterior, se debe ordenar el conjunto de elementos utilizando el valor absoluto de dichas diferencias en forma creciente (es decir, de menor a mayor sin considerar el signo). Una vez ordenado, se debe asignar a cada elemento un valor de posición o "rango" de la diferencia. Si hubiera dos o más diferencias que tengan el mismo valor absoluto, se debe asignar la media de los rangos implicados en el empate.

  3) A cada valor de "rango" obtenido, se le debe agregar el signo correspondiente a la diferencia de la que provino. En otras palabras, si la diferencia fue negativa, el rango pasa a ser negativo; en caso contrario, el signo del rango no cambia.

  4) Calcular, por una parte, la suma de los valores absolutos de los rangos negativos ($W-$) y, por la otra, la suma de los rangos positivos ($W+$).

5) Asignar *W* la menor suma de los dos rangos calculadas en el paso 4 ($W = min\{W-; W+\}$). Aunque se podría utilizar cualquiera de las dos sumas, este procedimiento selecciona arbitrariamente la más pequeña de las dos para ser más simple.

6) Determinar la cantidad de elementos o pares (*n*) como el número de pares de datos para los cuales la diferencia calculada en el paso 1 es diferente a cero. Con esta cantidad de pares y el nivel de significancia (α) previamente seleccionado, se debe buscar en la tabla F.1 el "valor crítico" correspondiente.

n	Valor Crítico según el nivel de significancia seleccionado (α)		
	$\alpha = 0,05$	$\alpha = 0,02$	$\alpha = 0,01$
6	0	-	-
7	2	-	-
8	3	1	-
9	5	3	1
10	8	5	3
11	10	7	5
12	13	9	7
13	17	12	9
14	21	15	12
15	25	19	15
16	29	23	19
17	34	27	23
18	40	32	27
19	46	37	32
20	52	43	37
21	58	49	42
22	65	55	48
23	73	62	54
24	81	69	61
25	89	76	68
26	98	84	75
27	107	92	83
28	116	101	91
29	126	110	100
30	137	120	109
31	147	130	118
32	159	140	128
33	170	151	138
34	182	162	148
35	195	173	159
36	208	185	171
37	221	198	182
38	235	211	194
39	249	224	207
40	264	238	220

Tabla F.1. Tabla de valores críticos para la prueba de rangos con signo de Wilcoxon.

7) Comparar la menor suma de los dos rangos calculados (*W*) con el valor crítico obtenido en el paso anterior:
    - Si W es menor o igual al valor crítico, entonces se rechaza la hipótesis nula (H_0).
    - Si W es mayor al valor crítico, no se rechaza la hipótesis nula (H_0).