Modelos para Asistir la Gestión de Proyectos de Explotación de Información

Pablo Pytel
Grupos Investigación en Sistemas de Información. DDPyT. Universidad Nacional de Lanús & GEMIS UTN-FRBA. Argentina.
ppytel@gmail.com

Paola Britos
Grupo de Investigación en Explotación de Información. Laboratorio de Informática Aplicada. Universidad Nacional de Río Negro. Argentina.
paobritos@gmail.com

Ramón García-Martínez
Grupo Investigación en Sistemas de Información. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús. Argentina.
rgarcia@unla.edu.ar

***Resumen*—Los proyectos de Explotación de Información son un tipo especial de proyecto de Ingeniería en Software. En lugar de requerir desarrollar un software específico, herramientas disponibles son utilizadas que ya incluyen las técnicas y algoritmos necesarios. Pero de todas formas posee problemas similares al existir cuestiones de gestión que todavía deben ser mejorados. Entre estas cuestiones se destaca la necesidad de un análisis de viabilidad que permita identificar los riesgos en forma temprana y un método de estimación de esfuerzo que asista la planificación de actividades y recursos que son necesarios para desarrollar el proyecto. En este contexto, este trabajo tiene como objetivo proponer y estudiar dos modelos para ser utilizados en Pequeñas y Medianas Empresas al comienzo de un proyecto de Explotación de Información y de esta forma buscar reducir los problemas que se pueden presentar.**

***Términos Clave* —Estimación, Viabilidad, PyMES, Explotación de Información.**

I. Introducción

La Explotación de Información consiste en la extracción de conocimiento no-trivial que reside de manera implícita en los datos disponibles en distintas fuentes de información [1]. Dicho conocimiento es previamente desconocido y puede resultar útil para algún proceso [2]. Una vez identificado el problema de Inteligencia de Negocio es posible definir el Proceso de Explotación de Información. Ese proceso se encuentra formado por varias técnicas de Minería de Datos que se ejecutan para lograr, a partir de un conjunto de información con un grado de valor para la organización, otro conjunto de información con un grado de valor mayor que el inicial [3]. Si bien existen metodologías que acompañan el desarrollo de proyectos de explotación de información que se consideran probadas y tienen un buen nivel de madurez entre las cuales se destacan CRISP-DM [4], P3TQ [5] y SEMMA [6], estas metodologías dejan de lado aspectos a nivel operativo de los proyectos y de empresa [7]. En estas metodologías se observa la falta de procesos y herramientas que permitan soportar de las actividades de gestión al inicio del mismo. Estas actividades son de gran importancia para reducir la probabilidad de fracasos en estos proyectos.

En este contexto, este trabajo tiene como objetivo proponer y estudiar dos modelos para ser utilizados en Pequeñas y Medianas Empresas (PyMEs) al comienzo de un proyecto de Explotación de Información y de esta forma buscar reducir los problemas que se pueden presentar. El primer modelo propuesto permite realizar la Evaluación de la Viabilidad del proyecto para así determinar así los posibles puntos fuertes y débiles. Por otro lado, el segundo modelo permite realizar la Estimación de los Recursos y Tiempo que serán requeridos para realizar el proyecto en forma satisfactoria.

Este trabajo incluye la siguiente estructura: primero se realiza una reseña de sobre los motivos por los que los proyectos pueden fracasar (sección II). Luego se identifican las principales características de estos proyectos para así proponer ambos modelos (sección III). Una vez que ambos modelos son propuestos, se presentan los resultados de su estudio con una prueba de concepto (sección IV) y una validación de casos (sección V). Finalmente, se indican las conclusiones obtenidas (sección VI).

II. Fracasos de Proyectos

La mayoría de los proyectos de Ingeniería en Software pueden ser considerados (al menos) fracasos parciales debido a que pocos proyectos cumplen con sus presupuestos de costo, planificación, criterios de calidad o especificaciones de requerimientos [8]. Para los proyectos cancelados o cuestionados, el proyecto promedio estuvo un 189% sobre el presupuesto, 222% retrasado en su planificación, y contenía sólo el 61% de las características originalmente solicitadas. En 2005 se consideraba que entre el 5 y 15% de los proyectos fueron abandonados antes o un poco después de la entrega por considerar totalmente inadecuados [9]. Según [10], entre los principales motivos que generan el fracaso de los proyectos se encuentra una pobre planificación, recursos insuficientes y falta de identificación de los riesgos.

Los proyectos de Explotación de Información son un tipo especial de proyecto de Ingeniería en Software. En lugar de requerir desarrollar un software específico, herramientas disponibles son utilizadas que ya incluyen las técnicas y algoritmos necesarios [3]. Como resultado las características de los proyectos de Explotación de Información son diferentes a los de la Ingeniería en Software Tradicional y de la Ingeniería del Conocimiento. Pero de todas formas posee problemas similares. Estudios realizados sobre sobre proyectos de Explotación de Información han detectado que la mayoría de los proyectos finaliza en fracaso por lo que no son terminados con éxito [11; 12]. En el año 2000 se ha había determinado que el 85% de los proyectos no alcanzan sus metas [13], mientras que en el 2005 el porcentaje de fracaso bajo a aproximadamente el 60% [14]. Por lo tanto se puede decir que la comunidad ha estado trabajando en el camino correcto pero

hay cuestiones de gestión que todavía deben ser mejorados. Entre estas cuestiones se destaca la necesidad de un análisis de viabilidad que permita identificar los riesgos en forma temprana y un método de estimación de esfuerzo que asista la planificación de actividades y recursos que son necesarios para desarrollar el proyecto.

III. MODELOS PROPUESTOS

En esta sección los dos modelos son propuestos para ser utilizados al comienzo de un proyecto de explotación de información. El primer modelo busca evaluar la viabilidad del proyecto (descripto en la subsección A) mientras que el segundo permite estimar los recursos y tiempo requerido (en meses/hombre) para desarrollar el proyecto (subsección B).

Tanto la definición como su posterior validación (realizada en el capítulo V) han utilizado información de proyectos reales de explotación de información que fueron recolectados 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).

Debe notarse que todos estos proyectos fueron realizados utilizando la metodología CRISP-DM [4], por lo que el método propuesto se considera confiable para proyectos de explotación de información a ser desarrollados con dicha metodología.

A. 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 el Modelo de Viabilidad, primero es necesario identificar las principales características que un Proyecto de Explotación de Información debe cumplir para ser considerado viable. El ingeniero encargado del proyecto deberá responder a dichas condiciones de acuerdo a las características del proyecto para así evaluar su viabilidad. Estas características han sido clasificadas en tres grupos:

- *Plausibilidad* que indica si es posible realizar el proyecto,
- *Adecuación* que determinar si la explotación de información es la mejor solución para el problema planteado, y
- *Éxito* que determinar si los resultados pueden y serán utilizados o no por la organización.

Sin embargo, como normalmente no es sencillo contestar las condiciones con respuestas del tipo 'sí' / 'no' (o dando una valoración numérica), el modelo propuesto deberá poder manejar un rango de valores lingüísticos en forma similar al criterio empleado por el test de viabilidad para proyectos de INCO indicado en [15; 16] que se encuentra basado en sistemas expertos difusos [17].

A partir de estos valores y aplicando un conjunto de pasos, se podrá obtener la valoración por dimensión y global de la viabilidad del proyecto. A continuación se describen los cinco pasos que se deben aplicar:

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 I las cuales fueron identificadas a partir de la investigación documental realizada en [18-25]. El ingeniero deberá responder las preguntas indicadas a partir del resultado de las entrevistas realizadas en la organización, asociadas a cada característica. Para ello, los valores lingüísticos permitidos 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.

TABLA I. CARACTERÍSTICAS EVALUADAS POR EL MODELO DE VIABILIDAD

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
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

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

- Categoría que se utiliza únicamente para poder agrupar las características de acuerdo a qué o quién se refiere.
- 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 a evaluar del proyecto.
- Peso que indica la importancia relativa a cada característica en la globalidad del modelo. Nótese que la suma de todos los pesos no es igual a 100 pero esto es soportado por las fórmulas utilizadas en el modelo.
- Umbral que indica el valor que la característica debe igualar o superar. En caso de que no supere el umbral, se puede considerar que el proyecto no es viable y no es necesario continuar con los siguientes pasos.

Paso 2: Convertir los valores en intervalos difusos.

Una vez que los valores lingüísticos han sido definidos para cada característica de la tabla 1, se deben traducir en

intervalos difusos. Para cada valor lingüístico se define un intervalo 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 1.

Paso 3: Calcular la valoración de cada dimensión.

Para calcular la valoración de cada dimensión del proyecto, los intervalos difusos (obtenidos en el paso anterior) son ponderados considerando su peso correspondiente (definido en la tabla 1). El intervalo que representa la valoración de cada dimensión (I_d) se calcula con la fórmula (1):

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

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*.

Está fórmula está 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.

Ya 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 fórmula (2):

$$V_d = \frac{\sum_{i=1}^{4} I_{d_i}}{4} \quad (2)$$

Donde:

- V_d: representa el valor numérico calculado para la dimensión *d*.
- I_{di}: representa el valor de la posición *i* del intervalo difuso calculado para la dimensión *d* .

Paso 4: 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 (la cual es indicada en la fórmula 3) y así se consigue el valor de la viabilidad global del proyecto (EV).

$$EV = \frac{8 \cdot V_P + 8 \cdot V_A + 6 \cdot V_E}{22} \quad (3)$$

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.

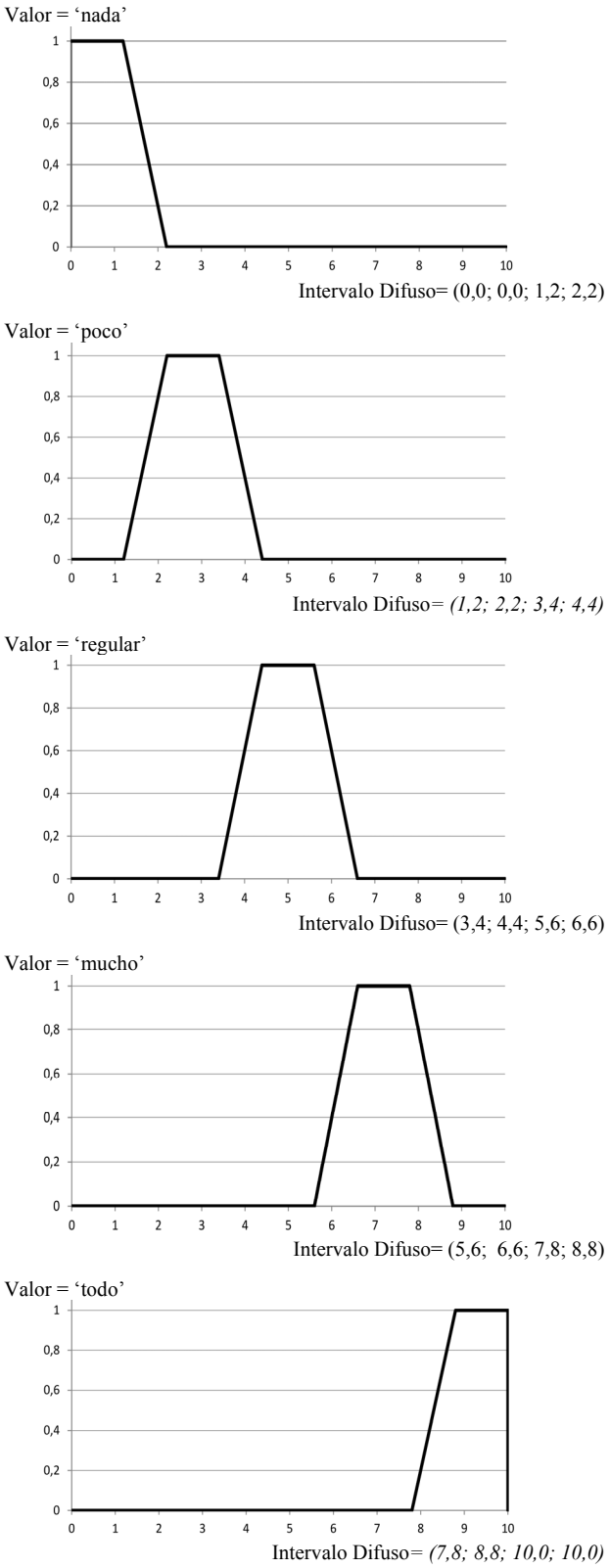


Fig. 1. Representación de la Función de Pertenencia y asignación de Intervalo Difuso para los Valores Lingüísticos.

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. Por un lado, para interpretar los resultados de la viabilidad de cada dimensión, se recomienda graficar la función de pertenencia del intervalo difuso (I_d). Se considera que la

viabilidad de la dimensión está aceptada si supera al intervalo del valor 'regular' (esto es análogo a considerar que el valor de la dimensión V_d es mayor a 5). Por otro lado, para la viabilidad del proyecto se utiliza el siguiente criterio: si las tres dimensiones son aceptadas (es decir el valor de cada dimensión es mayor a 5) 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 también podrá observar los puntos débiles del proyecto que debe reforzar (en caso de proyecto no viable) y/o monitorear durante el desarrollo del proyecto.

B. Modelo para la Estimación de Esfuerzo

Para realizar tanto una planificación como un presupuesto correcto para el proyecto se necesita una buena estimación al comienzo del mismo. De esta manera se destaca la necesidad de contar con métodos de estimación de esfuerzo confiables para proyectos de Explotación de Información. Dada las diferencias que existen entre un proyecto tradicional de construcción de software, los métodos usuales de estimación no son aplicables ya que los parámetros a ser utilizados son de naturalezas diferentes. No obstante en [26] se ha definido el modelo de estimación DMCoMo teniendo en cuenta las características particulares de los proyectos de Explotación de Información, un estudio detallado de DMCoMo ha demostrado que sólo es válido para proyectos grandes [27]. Al trabajar con proyectos medianos y pequeños DMCoMo tiende a sobreestimar el esfuerzo por lo que cual no es útil.

Teniendo en cuenta lo anterior, el modelo propuesto para la estimación considera las características de proyectos medianos y pequeños [28; 29] y las características de las Pequeñas y Medianas Empresas en Latinoamérica [30; 31] se ha propuesto un modelo que utiliza ocho factores de costos.

En esta propuesta se han definido pocos factores de costo, ya que como se demuestra en [33], al momento de crear un nuevo método de estimación es preferible ignorar muchos de los datos no significativos para evitar que el modelo sea demasiado complejo y por lo tanto poco práctico. De esta manera se busca eliminar las variables tanto irrelevantes como dependientes, y además reducir la varianza y el ruido. Los factores de costo han sido seleccionados teniendo en cuanta las tareas más críticas de la metodología CRISP-DM [4]: en [33] se indica que actualmente la construcción de los modelos de minería de datos y buscar los 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, 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 los goles 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 proyecto:

•Tipo de objetivo de explotación de información (OBTY)

Este factor de costo analiza el objetivo del proyecto de Explotación de Información considerando el tipo de proceso a ser aplicado para obtenerlo de acuerdo a la definición realizada en [3] de acuerdo a los datos disponibles y las metas del proyecto. Los posibles valores de este factor de costo se indican en la tabla II.

TABLA II. VALORES DEL FACTOR DE COSTO OBTY

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.

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

El grado de apoyo y participación de los miembros de la organización se analiza viendo si la alta gerencia (normalmente los dueños de la PyME), la gerencia media (supervisores y/o jefes de área) y/o el resto del personal están dispuestos a ayudar al equipo de trabajo para comprender el negocio y los datos. Se sobreentiende que si un proyecto de explotación de información fue contratado, por lo menos la alta gerencia va a apoyar el mismo. Los posibles valores de este factor de costo se indican en la tabla III.

TABLA III. VALORES DEL FACTOR DE COSTO LECO

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.

Factores de costo relacionados a los datos disponibles:

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

Se analizan los repositorios 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 disponibles (públicos o privados de la organización) como la tecnología en que se encuentran implementadas. No interesa conocer la cantidad de tablas que posee cada repositorio dado que se entiende que la integración de los datos dentro de un repositorio es relativamente sencilla (sobre todo al utilizar sistemas gestores de bases de datos por poder ser realizada con un comando query). Sin embargo, dependiendo de la tecnología, la complejidad de las tareas de integración de los datos puede ser mayor o menor.

Los criterios recomendados para ser utilizados se describen a continuación:

- 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 al realizar la integración con estos datos exportados.

- Por otro lado, si existen repositorios que no están en forma digital (es decir impreso en papel) se considera que la tecnología será no compatible pero 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 posibles valores de este factor de costo se indica en la tabla IV.

TABLA IV. VALORES DEL FACTOR DE COSTO AREP

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.

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

Este factor de costo considera la cantidad total de tuplas (registros) disponibles en la tabla principal utilizada para aplicar los algoritmos de minería de datos. Los posibles valores de este factor de costo se indican en la tabla V.

TABLA V. VALORES DEL FACTOR DE COSTO QTUM

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.

• Cantidad de tuplas disponibles en tablas auxiliares (QTUA)

Esta variable considera la cantidad aproximada de tuplas (registros) disponibles en las tablas auxiliares (si existieran) que son utilizadas para agregar información a la tabla principal (como es la tabla que define las características del producto a partir de su identificador en la tabla de ventas). Estas tablas auxiliares normalmente suelen tener menos registros que la tabla principal. Los posibles valores de este factor de costo se indican en la tabla VI.

TABLA VI. VALORES DEL FACTOR DE COSTO QTUA

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.

• Nivel de conocimiento sobre los datos (KLDS)

Considera el nivel de documentación existente sobre los repositorios de datos. Es decir, se analiza si existe un documento donde se indique la tecnología en que están implementados, los campos 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, será necesario realizar reuniones con los expertos (normalmente los encargados de la administración y mantenimiento de los repositorios). Los posibles valores de este factor de costo se indican en la tabla VII.

TABLA VII. VALORES DEL FACTOR DE COSTO KLDS

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.

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 ocupará de llevar adelante el proyecto. 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 similares en el mismo tipo de negocio y los datos a ser utilizados.

Por lo tanto se debe evaluar el conocimiento y experiencia previa en proyectos anteriores similares al que se está llevando a cabo con respecto al tipo de negocio, los da-tos a ser utilizados y los objetivos que se esperan lograr. Los posibles valores de este factor de costo se indican en la tabla VIII.

TABLA VIII. VALORES DEL FACTOR DE COSTO KEXT

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.

• Funcionalidad de las herramientas disponibles (TOOL)

Finamente, este factor de costo evalúa las características de las herramientas disponibles para realizar el proyecto. Para ello se analiza tanto las funcionalidades de preparación de los datos como los algoritmos de minería de datos que posee implementadas. Sus posibles valores de este factor de costo se indican en la tabla IX.

Una vez que los factores de costo fueron definidos, se han utilizado para caracterizar 34 proyectos reales de explotación de información recolectados los cuales se encuentran disponibles en [34]. Estos proyectos provistos por colegas investigadores (como se ha indicado anteriormente) incluyen el esfuerzo real que fue requerido para realizar el proyecto en forma completa.

TABLA IX. VALORES DEL FACTOR DE COSTO TOOL

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.

Una vez que los factores de costo fueron definidos, se han utilizado para caracterizar 34 proyectos reales de explotación de información recolectados los cuales se encuentran disponibles en [34]. Estos proyectos provistos por colegas investigadores (como se ha indicado anteriormente) incluyen el esfuerzo real que fue requerido para realizar el proyecto en forma completa.

Un método de regresión lineal multivariante [35] fue aplicado sobre estos datos para obtener una ecuación lineal de la forma utilizada por los métodos de la familia COCOMO [36]. Como resultado del proceso de regresión se obtiene la fórmula (4) que se indica a continuación.

$$\begin{aligned} PEM = {} & 0{,}80 \cdot OBTY + 1{,}10 \cdot LECO \\ & - 1{,}20 \cdot AREP - 0{,}30 \cdot QTUM \\ & - 0{,}70 \cdot QTUA + 1{,}80 \cdot KLDS \\ & - 0{,}90 \cdot KEXT + 1{,}86 \cdot TOOL \\ & - 3{,}30 \end{aligned} \qquad (4)$$

Donde:

PEM es el esfuerzo estimado por el método de estimación para PyMEs (en meses/hombre)

OBTY, LECO, AREP, QTUM, QTUA, KLDS, KEXT y TOOL son los valores correspondientes de los factores de costo definidos en las tablas II a IX respectivamente.

IV. PRUEBA DE CONCEPTO

A modo de ejemplo en esta sección se presenta una prueba de concepto de los modelos propuestos. Para ello se utilizan los datos de un proyecto de explotación de información real finalizado con éxito (y por lo tanto fue viable) que fue desarrollado por 3 personas en 4 meses (es decir que tuvo un esfuerzo de 12 meses/hombre o un año/hombre).

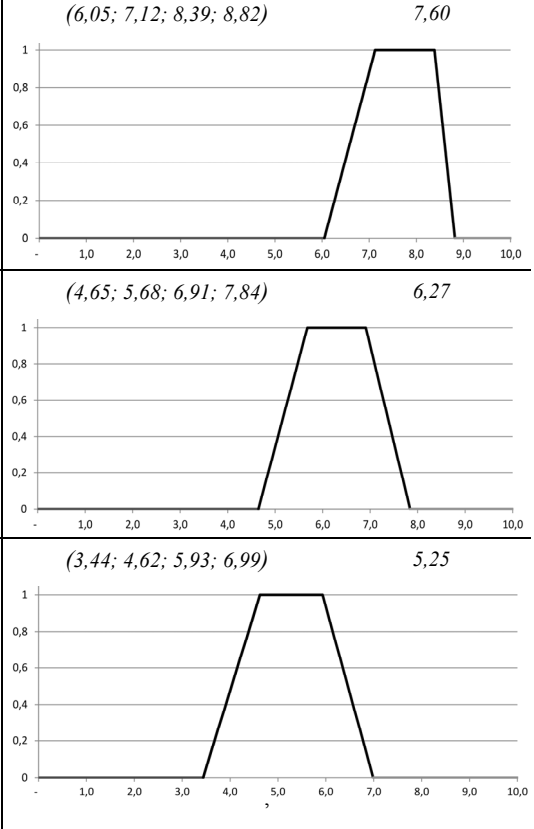
Este proyecto tenía el objetivo de detectar las evidencias de causalidad entre la satisfacción general de los clientes de una organización proveedora de Internet mediante la detección de evidencias de causalidad entre satisfacción general, servicio contratado y baja de clientes. Para ello se utilizó la información de una encuesta realizada por la organización a sus clientes.

Para realizar la prueba de concepto, primero se aplicaron los cinco pasos propuestos en el Modelo para Evaluar la Viabilidad (en la sección III-A). Todos los cálculos necesarios fueron realizados mediante una planilla creada ad-hoc [37] que implementa las fórmulas definidas anteriormente y sus resultados se muestran en la tabla X.

Como se puede ver, dado que los valores de cada dimensión y de la viabilidad global del proyecto son superiores al valor mínimo requerido, se considera que el proyecto es viable para ser realizado (lo cual es correcto).

TABLA X. RESULTADOS DEL MODELO DE VIABILIDAD

Dimensión	Intervalo Valor Dimensión (I_d)	Valor de la Dimensión (V_d)
Plausibilidad	*(6,05; 7,12; 8,39; 8,82)*	*7,60*
Adecuación	*(4,65; 5,68; 6,91; 7,84)*	*6,27*
Éxito	*(3,44; 4,62; 5,93; 6,99)*	*5,25*
Valor global de la viabilidad del proyecto (*EV*)		***6,47***



Sin embargo, el proyecto posee algunos puntos débiles. Puede notarse que a pesar de que la valoración de *Plausibilidad* y *Adecuación* es holgada (valores cercanos a '*mucho'*), para el *Éxito* del proyecto es muy cercana al valor mínimo requerido (es decir al intervalo correspondiente al valor '*regular'*). Esto significa que se debe prestar mayor atención las características asociadas al éxito para asegurar que el proyecto se desarrolle sin problemas. De esta forma estos puntos débiles se convierten en riesgos para ser monitoreados y controlados durante el desarrollo del proyecto.

Por otro lado, ya que el proyecto se considera viable se utiliza el Modelo de Estimación de Esfuerzo (sección III-B) en este proyecto para comparar el esfuerzo real del proyecto con el calculado por el modelo.

Para ello se definen los valores de cada factor de costo y luego se aplica la fórmula correspondiente para obtener el esfuerzo. En la tabla XI, se detallan los valores de los factores de costo utilizados para la prueba de concepto.

Como resultado de aplicar la fórmula del modelo se obtiene un esfuerzo estimado de 12,18 meses/hombre. Al compararlo con el esfuerzo real que fue requerido por el proyecto de 12 meses/hombre se puede ver que el error del modelo es menor a 6 días/hombre (es decir 0,18 meses/hombre) con respecto al real. Esto significa que para este ejemplo el modelo fue muy preciso.

TABLA XI. RESULTADOS DEL MODELO DE ESTIMACIÓN

Categoría	ID	Descripción	Valor
Proyecto	OBTY	*Se desea conocer la incidencia de los atributos sobre el motivo de baja del servicio.*	*5*
	LECO	*Se posee buena disposición del personal y la dirección para colaborar en el proyecto.*	*1*
Datos Disponibles	AREP	*Sólo 1 repositorio disponible.*	*1*
	QTUM	*Aprox. 15.000 tuplas en la tabla principal.*	*3*
	QTUA	*Aprox. 10.000 tuplas en tablas auxiliares*	*3*
	KLDS	*Las tablas y repositorios no están documentados y no existen expertos.*	*6*
Recursos Disponibles	KEXT	*Se ha trabajado en tipos de organizaciones similares pero con datos diferentes para mismos objetivos.*	*2*
	TOOL	*La herramienta posee funciones para el formateo y para aplicar las técnicas de minería de datos, pero sólo importa una tabla.*	*3*
Esfuerzo Estimado por el Modelo = 12,18 meses/hombre			

V. VALIDACIÓN DE LOS MODELOS PROPUESTOS

En esta sección se presentan los resultados de la validación realizada sobre los modelos propuestos en la sección III. Para realizar esta validación se han utilizado datos de 25 proyectos reales, cuyos datos proyectos se encuentran disponibles en [38]. En ese reporte se incluye también todos los cálculos auxiliares utilizados por el Modelo de Viabilidad. Un resumen de estos datos se muestra en la tabla XII. Como se puede ver veintidós de los proyectos (P1 a P22 inclusive) han finalizado con éxito y los tres restantes (P23, P24 y P25) fueron cancelados.

Para la validación de los modelos primero se realiza el análisis de gráficos Boxplot y luego en aplicar la prueba de rangos con signo de Wilcoxon [39]. Esta prueba no paramétrica se utiliza para comprobar que no hay diferencia entre los datos reales y los calculados por cada modelo.

A. Validación del Modelo para la Evaluación de la Viabilidad

Como se puede ver en la tabla XII, se les ha pedido a los investigadores que indiquen una valoración de las cuatro dimensiones consideradas por el modelo (plausibilidad, adecuación y éxito) utilizando una escala de 1 a 10 (donde 10 es el mayor valor). Con estos valores se calcula su promedio para obtener la Valoración de la Viabilidad del proyecto. Esta valoración se utiliza junto con los resultados de aplicar el procedimiento propuesto para validar el modelo propuesto.

Para realizar las pruebas se toma cada dimensión (plausibilidad, adecuación, éxito y viabilidad global) en forma independiente. Primero se compara el comportamiento del modelo con la valoración de los investigadores en gráficos boxplot (figura 2). En estos gráficos se muestran los valores mínimos y máximo, el rango del desvío estándar y el valor medio para cada uno.

TABLA XII. DATOS REALES DE CADA PROYECTO Y LOS CALCULADOS POR LOS MODELOS

#	Valoración indicada por los Investigadores				Esfuerzo Real*	Valores calculados por el Modelo de Viabilidad				Esfuerzo calculado por el Modelo de Estimación*
	Plausibilidad	Adecuación	Éxito	Viabilidad Global		Plausibilidad	Adecuación	Éxito	Viabilidad Global	
P1	*8*	*7*	*4*	***6,33***	***2,41***	*7,20*	*6,11*	*5,25*	***6,27***	***2,58***
P2	*7*	*6*	*5*	***6,00***	***7,00***	*6,87*	*5,07*	*5,25*	***5,77***	***6,00***
P3	*8*	*5*	*6*	***6,33***	***1,64***	*5,90*	*5,67*	*5,31*	***5,65***	***1,48***
P4	*6*	*6*	*4*	***5,33***	***3,65***	*5,12*	*6,95*	*4,12*	***5,51***	***1,68***
P5	*6*	*8*	*7*	***7,00***	***9,35***	*5,12*	*7,82*	*6,81*	***6,56***	***9,80***
P6	*6*	*5*	*5*	***5,33***	***11,63***	*5,45*	*5,61*	*5,25*	***5,45***	***5,10***
P7	*5*	*5*	*5*	***5,00***	***6,73***	*5,45*	*5,56*	*5,42*	***5,48***	***3,78***
P8	*6*	*5*	*6*	***5,67***	***5,40***	*6,45*	*5,80*	*5,18*	***5,87***	***4,88***
P9	*7*	*6*	*6*	***6,33***	***8,38***	*7,20*	*5,61*	*5,57*	***6,18***	***8,70***
P10	*6*	*5*	*6*	***5,67***	***1,56***	*5,85*	*5,34*	*5,57*	***5,59***	***1,08***
P11	*8*	*5*	*6*	***6,33***	***9,70***	*6,22*	*6,56*	*5,42*	***6,14***	***9,60***
P12	*7*	*8*	*7*	***7,33***	***5,24***	*7,67*	*7,35*	*6,45*	***7,22***	***5,80***
P13	*7*	*5*	*6*	***6,00***	***5,00***	*5,93*	*5,09*	*7,05*	***5,93***	***4,58***
P14	*7*	*7*	*6*	***6,67***	***8,97***	*6,20*	*6,59*	*5,69*	***6,20***	***9,18***
P15	*9*	*7*	*8*	***8,00***	***2,81***	*8,72*	*6,89*	*7,66*	***7,77***	***3,48***
P16	*7*	*6*	*5*	***6,00***	***11,80***	*6,45*	*6,43*	*5,64*	***6,22***	***12,00***
P17	*6*	*5*	*5*	***5,33***	***2,79***	*6,14*	*5,83*	*5,42*	***5,83***	***2,28***
P18	*5*	*5*	*6*	***5,33***	***3,88***	*6,00*	*5,31*	*5,42*	***5,59***	***3,58***
P19	*8*	*7*	*7*	***7,33***	***5,70***	*7,01*	*6,89*	*5,58*	***6,58***	***6,30***
P20	*9*	*7*	*5*	***7,00***	***8,54***	*8,24*	*6,75*	*5,52*	***6,96***	***9,18***
P21	*8*	*6*	*5*	***6,33***	***10,61***	*8,05*	*6,45*	*5,25*	***6,70***	***11,50***
P22	*7*	*6*	*6*	***6,33***	***6,88***	*6,45*	*5,81*	*6,54*	***6,24***	***6,40***
P23	*3*	*4*	*3*	***3,33***	-	*4,66*	*5,34*	*3,25*	***4,52***	-
P24	*5*	*3*	*2*	***3,33***	-	*4,66*	*3,46*	*4,21*	***4,10***	-
P25	*4*	*2*	*1*	***2,33***	-	*4,63*	*2,81*	*3,01*	***3,52***	-

**: en meses/hombre y sólo indicado para proyectos que han finalizado con éxito.*

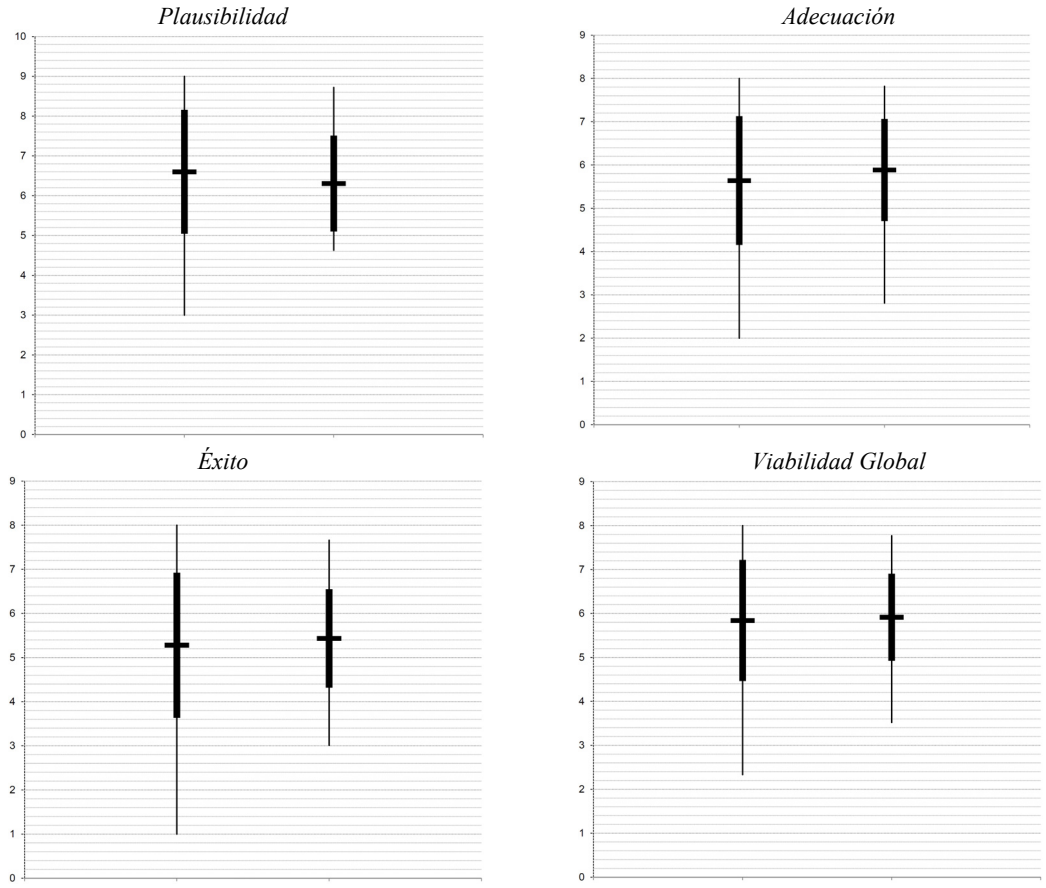


Fig. 2. Gráficos boxplot para el Modelo de Viabilidad

Como se puede ver el comportamiento para las cuatro dimensiones es muy similar por ser tanto los valores medio como el rango del desvío estándar muy similares (la mayor diferencia es de 0,3 para la plausibilidad). Sin embargo, el modelo propuesto tiende a ser más conservador por no tener valores tan extremos sobre todo para el mínimo.

Dado que las diferencias de los pares de datos tienen una distribución que es aproximadamente simétrica se puede aplicar el análisis de la prueba de rangos con signo de Wilcoxon. En esta prueba se aplican la siguiente hipótesis nula (H_0) y alternativa (H_1):

H_0: *Los valores indicados por los investigadores y 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 definido 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 definido por el modelo).*

Los resultados de aplicar la prueba de Wilcoxon se muestran en la tabla XIII.

Como en todos los casos se obtuvo 25 pares o proyectos con diferencias distinta de cero (*n=25*) y el nivel de significancia seleccionado es de 0,01 entonces la hipótesis nula (H_0) será rechazada si la menor suma de rangos (W) es menor o igual a 68 (valor crítico obtenido de la tabla estadística). En caso contrario, no se rechaza y se considera como válida. En el caso de la *Plausibilidad* se toma 97 como la menor suma de rangos (W) por ser la suma de rangos positiva (W^+) la menor. Dado que este valor es mayor que el valor crítico (68), no se rechaza la hipótesis nula y se puede decir que no hay diferencia entre el valor calculado por el modelo para la plausibilidad y el asignado por los investigadores. Para la *Adecuación*, como W^- es la menor, se toma W = 98. Dado que este valor es mayor que 68, no se rechaza la hipótesis nula (H_0). Con el *Éxito* sucede algo similar: $W = W^- = 150 > 68$, entonces no se rechaza H_0. Finalmente la *Viabilidad Global* tampoco rechaza H_0 ($W = W^- = 144 > 68$).

TABLA XIII. RESULTADOS DE APLICAR LA PRUEBA DE WILCOXON AL MODELO DE VIABILIDAD

Dimensión	Suma de Rangos$^+$ (W^+)	Suma de Rangos$^-$ (W^+)	Cantidad de pares distintos de cero
Plausibilidad	*97*	*228*	*25*
Adecuación	*227*	*98*	*25*
Éxito	*175*	*150*	*25*
Viabilidad Global	*181*	*144*	*25*

Por lo tanto, con un se puede decir con un nivel de significancia del 0,01 que no hay diferencia entre el valor calculado por el modelo para todas las dimensiones y el asignado en la valoración realizada por los investigadores.

B. Validación del Modelo para la Estimación de Esfuerzo

Para analizar el comportamiento del modelo de estimación orientado para PyMEs propuesto se utiliza sólo la información de los primeros 22 proyectos indicados en la tabla XII (P1 a P22 inclusive) por haber finalizado exitosamente y ser conocido su esfuerzo real. Con estos proyectos se ha calculado los esfuerzos por el modelo propuesto con la fórmula PEM definida en la sección III-B.

Para comparar con mayor claridad el esfuerzo real con el calculado se genera un gráfico boxplot (figura 3). Se puede observar que se produce un error promedio de aproximadamente 0,49 meses/hombre con un desvío estándar para el error de aproximadamente ± 1,62 meses/hombre. Se nota que el modelo propuesto tiende a generar estimaciones un poco inferiores a las reales (el promedio del esfuerzo real es de 6,35 meses/hombre y la del modelo es de 5,86 meses/hombre) pero en la mayoría de los casos (19 proyectos de los 22 analizados) el error es menor al 35% del esfuerzo real.

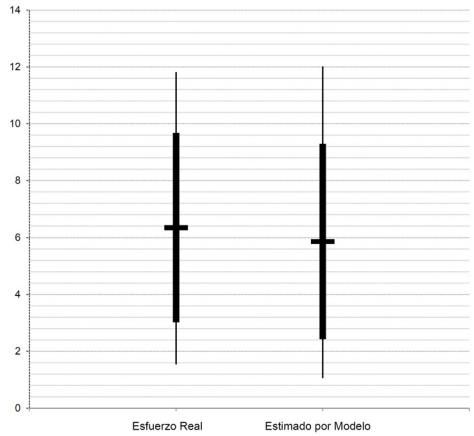


Fig. 3. Gráfico boxplot para el Modelo de Estimación

Dado que las diferencias de los pares de datos tienen una distribución que es aproximadamente simétrica también se puede aplicar el análisis de la prueba de rangos con signo de Wilcoxon. En esta nueva prueba se aplican la siguiente hipótesis nula y alternativa:

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 lo requerido realmente y lo definido 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 lo definido por el modelo).*

Los resultados de aplicar la prueba de Wilcoxon se muestran a continuación:

- Suma de Rangos$^+$ (W^+) = *108*
- Suma de Rangos$^-$ (W^-) = *145*

Como en todos los casos se obtuvo 22 pares con diferencias distinta de cero (*n=22*) y el nivel de significancia seleccionado es de 0,01 entonces el valor crítico obtenido de la tabla estadística es 49. Dado que la mínimo menor suma de rangos (W) es igual a 108 (W^+) y es mayor a 49, 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 el modelo propuesto.

VI. Conclusión

Se ha observado en los Proyectos de Explotación de Información la ausencia de modelos que permitan identificar sus riesgos al inicio del mismo y estimar el esfuerzo requerido cuando se desarrollan en PyMEs. Esto genera que de la gran cantidad de proyectos desarrollados, no todos finalizan con éxito, terminando la mayoría en fracasos. Por lo tanto, en este trabajo incluye dos propuestas:

- Primero se define un Modelo que permita la Evaluación de la Viabilidad para Proyectos de Explotación de Información usando la información disponible al comienzo del mismo. Mediante un procedimiento que consta de cinco pasos es posible caracterizar un proyecto y calcular la viabilidad de acuerdo a tres dimensiones: plausibilidad, adecuación y éxito. Esta evaluación además permite identificar los puntos débiles del proyecto. A pesar de que el proyecto sea viable, estos puntos débiles deben ser monitoreados durante el desarrollo del proyecto como riesgos. Es responsabilidad del ingeniero mantener o "subir" su valor para evitar así el fracaso del proyecto.
- Además se propone un Modelo que permite Estimar el Esfuerzo que se necesita para realizar el proyecto en su totalidad. Debe notarse que este modelo se encuentra orientado a las particularidades de los proyectos de corto alcance que son usualmente requeridos por las Pequeñas y Medianas Empresas. Para ello, se vuelve a caracterizar el proyecto esta vez mediante la asignación de valores a un conjunto de factores de costos que luego se utilizan para calcular el esfuerzo mediante una fórmula similar a las utilizadas por los métodos de la familia COCOMO.

Para ambos modelos se realiza una validación mediante su aplicación en proyectos reales y su comparación. Como resultado se determina que ambos modelos producen resultados muy precisos. En ambos casos el comportamiento general de los modelos tiende a ser similar al de los proyectos reales.

AGRADECIMIENTOS

Este trabajo de investigación ha si parcialmente financiado por los proyectos 33A167 y 33B102 de la Universidad Nacional de Lanús, por los proyectos 40B133 y 40B065 de la Universidad Nacional de Río Negro, y el proyecto EIUTIBA11211 de la Universidad Tecnológica Nacional Facultad Regional Buenos Aires. Además los autores desean agradecer a los investigadores que han provisto la información de proyectos reales utilizados.

REFERENCIAS

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

[2] Stefanovic, N., Majstorovic. V., & Stefanovic, D. "Supply Chain Business Intelligence Model". Proceedings 13th International Conference on Life Cycle Engineering. 2006. pp. 613-618.

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

[4] Chapman, P., Clinton, J., Keber, R., et al.. "CRISP-DM 1.0 Step by step BI guide". Edited by SPSS. 2000. http://tinyurl.com/crispDM

[5] Pyle, D. Business "Modeling and Business intelligence". Morgan Kaufmann. 2003.

[6] SAS Enterprise Miner: "SEMMA". 2008. http://tinyurl.com/semmaSAS

[7] Vanrell, J., Bertone, R., & García-Martínez, R. "Modelo de Proceso de Operación para Proyectos de Explotación de

Información". Anales del XVI Congreso Argentino de Ciencias de la Computación, 674-682. ISBN 978-950-9474-49-9. 2010.

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

[9] Charette, R.N. "Why software fails", Spectrum, IEEE, 42(9), pp. 42-49. 2005.

[10] The Standish Group: "Chaos Report 2010". http://blog.standish group. com/

[11] Edelstein, H.A. & Edelstein, H.C., "Building, Using, and Managing the Data Warehouse", Data Warehousing Institute, Prentice-Hall PTR, EnglewoodCliffs (NJ). 1997.

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

[13] Fayyad, U.M. "Tutorial report". Summer school of DM. Monash University (Australia). 2000.

[14] Gondar, J.E. "Metodología del Data Mining". Number 84-96272-21-4. Data Mining Institute S.L.. 2005.

[15] García-Martínez, R. & Britos, P. "Ingeniería de Sistemas Expertos". Editorial Nueva Librería. 2004. ISBN 987-1104-15-4.

[16] Gómez, A., Juristo, N., Montes, C. & Pazos, J. "Ingeniería del Conocimiento", Ed. Ramón Areces S.A. (Madrid). 1997.

[17] Jang, J.S.R. "Fuzzy inference systems", Upper Saddle River, NJ: Prentice-Hall. 1997.

[18] Sim, J. "Critical success factors in data mining projects". Ph.D. Thesis, University of North Texas. 2003.

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

[20] Davenport, T.H. "Make Better Decisions", Harvard Business Review, (November), pp. 117-123. 2009.

[21] Bolea, U., Jakličb, J. Papac, G. & Žabkard, J. "Critical Success Factors of Data Mining in Organizations", Ljubljana. 2011.

[22] Nadali, A., Kakhky, E.N. & Nosratabadi, H.E. "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, IEEE Vol. 6, pp. 161-165. 2011. doi:10.1109/ICECTECH.2011.5942073.

[23] Nie, G., Zhang, L., Liu, Y. Zheng, X. & Shi, Y. "Decision analysis of data mining project based on Bayesian risk", Expert Systems with Applications, 36(3), pp. 4589-4594. 2009.

[24] Pipino, L.L., Lee, Y.W. & Wang, R.Y. "Data quality assessment", Communications of the ACM, 45(4), pp. 211-218. 2002.

[25] Lavravc, N., Motoda, H., Fawcett, T., Holte, R. Langley, P. & Adriaans, P. "Introduction: Lessons learned from data mining applications and collaborative problem solving", Machine learning, vol. 57, n.º 1, pp. 13-34. 2004.

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

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

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

[29] Laporte, C., Alexandre, S. & Renault, A. "Developing International Standards for VSEs". Computer, 41(3): 98. 2008.

[30] Organization for Economic Cooperation and Development. "OECD SME and Entrepreneurship Outlook 2005". OECD Publishing. 2005.

[31] Álvarez, M. & Durán, J. "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. 2009.

[32] Chen, Z., Menzies, T., Port, D., et al. "Finding the right data for software cost modeling". Software, IEEE, vol.22, no.6, pp. 38-46, Nov.-Dec. 2005.

[33] Domingos, P., Elkan, C., Gehrke, J., et al. "10 challenging problems in data mining research". International Journal of Information Technology & Decision Making, 5(4): 597. 2006.

[34] Pytel, P. "Datos Recopilados para Estimación de Proyectos de Explotación de Información en PYMEs", Reporte Técnico GISI-TD-2011-01-RT-2012-01, http://www.unla.edu.ar/sistemas/gisi/GISI/papers/GISI-TD-2011-01-TR-2012--DatosProyectos.pdf

[35] Weisberg, S. "Applied Linear Regression". John Wiley & Sons, New York. 1985.

[36] Boehm, B., Abts, C., Brown, A., Chulani, S., Clark, B., Horowitz, E., Madachy, R., Reifer, D., Steece, B. "Software Cost Estimation with COCOMO II". Prentice-Hall. 2000.

[37] Pytel, P. Implementación del Modelo de Viabilidad Propuesto. 2012. http://tinyurl.com/ViabPruConcepto

[38] Pytel, P. "Datos Recopilados para Validación del Modelo de Viabilidad de Proyectos de Explotación de Información", Reporte Técnico GISI-TD-2011-01-RT-2012-02, http://www.unla.edu.ar/sistemas/gisi/GISI/papers/GISI-TD-2011-01-TR-2012-02-20Datos-Proyectos-para-Viabilidad.pdf

[39] Wilcoxon, F. "Individual Comparisons by Ranking Methods", Biometrics 1, pp. 80-83. 1945.

Pablo Pytel. Es Ingeniero en Sistemas de Información por la UTN, Magister en Ingeniería de Software por la Universidad Politécnica de Madrid. Es Docente Instructor en la Licenciatura en Sistemas y codirector del proyecto UNLa 33B102 de la UNLa. Sus intereses en investigación son modelos de viabilidad y estimación de proyectos de explotación de información; y aplicaciones de IA a IS.

Paola Britos. Es Licenciada en Sistemas de Información por la UNLu, Magister en Ingeniería del Conocimiento por la Universidad Politécnica de Madrid y Doctora en Ciencias Infromáticas por la Universidad Nacional de La Plata. Es Profesora Asociada Regular del Área de Ingenieria de Software en la Licenciatura en Sistemas de Infromación y directora de los proyectos 40B133 y 40B065 de la UNRN. Sus intereses en investigación son modelos de proceso para proyectos de explotación de información.

Ramón García Martínez. Es Analista de Computación por la UNLP, es Licenciado en Sistemas de Información por la UNLu, es Master en Ingeniería Informática y Doctor en Informática por la Universidad Politécnica de Madrid. Es Profesor Titular Regular del Area de Ingeniería de Software en la Licenciatura en Sistemas y Director de los proyectos 33A166 y 33A167 la UNLa. Su áreas de interés en investigación son Aprendizaje Automático, Sistemas Inteligentes, Explotación de Datos basada en Sistemas Inteligentes, Ingeniería del Conocimiento y las correspondientes aplicaciones en Ingeniería, Economía, Salud y Agroindustria.