Propuesta de Métricas para Proyectos de Explotación de Información

Diego Basso[1,2], Darío Rodríguez[3], Ramón García-Martínez[3]

1. Grupo de Investigación en Ingeniería. Departamento de Ingeniería e Investigaciones Tecnológicas. Universidad Nacional de La Matanza
2. Programa de Maestría en Ingeniería de Sistemas de Información. UTN-FRBA
3. Laboratorio de Investigación y Desarrollo en Ingeniería de Explotación de Información. Grupo de Investigación en Sistemas de Información. Universidad Nacional de Lanús.
diebasso@yahoo.com.ar, drodrigu@unla.edu.ar, rgm1960@yahoo.com

Resumen. Los proyectos de explotación de información requieren de un proceso de planificación que permita estimar sus tiempos y medir el avance del producto en cada etapa de su desarrollo y calidad del mismo. Las métricas usuales para realizar una estimación no se consideran adecuadas ya que los parámetros a ser utilizados son de naturaleza diferentes y no se ajustan a sus características particulares. En este contexto, se plantea una propuesta de métricas aplicables al proceso de desarrollo de Proyectos de Explotación de Información, siguiendo los lineamientos del Modelo de Procesos para Proyectos de Explotación de Información para PyMEs.

Palabras claves: Métricas. Explotación de Información. Modelo de Procesos

1. Introducción

En la Ingeniería de Software, los proyectos de desarrollo tradicionales aplican una amplia diversidad de métricas e indicadores a distintos atributos y características de los productos y procesos del desarrollo, con el fin de garantizar la calidad del software construido. En el ámbito de la Ingeniería en Conocimiento, especialmente en el desarrollo de sistemas expertos o sistemas basados en conocimientos, mediante la medición de la conceptualización se puede estimar actividades futuras y obtener información del estado de madurez del conocimiento sobre el dominio y sus particularidades [Hauge et al., 2006]. Estas métricas de madurez de conceptualización para Sistemas Expertos aplicadas en [Pollo-Cattaneo, 2007] brindan además información sobre la complejidad del dominio. Los Proyectos de Explotación de Información también requieren de un proceso de planificación que permita estimar sus tiempos y medir el avance del producto en cada etapa de su desarrollo y calidad del mismo. Sin embargo, dada las diferencias que existen con un proyecto clásico de construcción de software, las métricas usuales para realizar una estimación no se consideran completamente adecuadas ya que los parámetros a ser utilizados son de naturaleza diferentes y no se ajustan a sus características particulares.

Este trabajo tiene como objetivo presentar una propuesta de métricas aplicables al proceso de desarrollo de proyectos de Explotación de Información propuesto en [Vanrell, 2011] para PyMEs. Para ello, primero se realiza una introducción a las características del proceso de desarrollo mencionado, los procesos de explotación de información basados en sistemas inteligentes y la categorización definida en el método DMCoMo para agrupar dichas características (sección 2); luego se delimita el problema (sección 3) presentando una propuesta de solución (sección 4) y las consideraciones para la validación de la propuesta (sección 5), finalizando con la puntualización de algunas conclusiones parciales (sección 6).

2. Estado de la Cuestión

En esta sección se introduce el proceso de desarrollo para proyectos de Explotación de Información (sección 2.1), los procesos de explotación de información basados en sistemas inteligentes (sección 2.2) y el método de estimación desarrollado para este tipo de proyectos - DMCoMo (sección 2.3).

2.1. Proceso de desarrollo para proyectos de Explotación de Información

Las etapas de desarrollo de los proyectos de Explotación de Información no coinciden naturalmente con las etapas mediante las cuales se desarrollan los proyectos de software tradicionales. El modelo de procesos para proyectos de Explotación de Información propuesto en [Vanrell, 2011] plantea dos procesos principales: uno vinculado a la administración de proyectos de explotación de información y otro relacionado con el desarrollo del proyecto. Para el interés de este trabajo, nos centraremos en el segundo de los procesos mencionados, cuyos subprocesos y tareas están definidas a partir de las fases de desarrollo planteadas por la metodología CRISP-DM. Estos subprocesos son: Entendimiento del Negocio, Entendimiento de los Datos, Preparación de los Datos, Modelado, Evaluación y Entrega. Claramente estos subprocesos difieren de las etapas definidas para un proyecto de desarrollo de software tradicional (inicio, requerimientos, análisis y diseño, construcción, integración y pruebas y cierre).

2.2. Procesos de Explotación de Información basados en sistemas inteligentes

En el trabajo realizado por [Britos, 2008] se proponen cinco procesos de explotación de información que podrían aplicarse a la etapa de Modelado del proceso de desarrollo propuesto por [Vanrell, 2011]. Estos procesos son:

- Proceso de Descubrimiento de Reglas: permite identificar condiciones para obtener resultados del dominio del problema.
- Proceso de Descubrimiento de Grupos: permite identificar una partición dentro de la información disponible dentro del dominio de un problema.

- Proceso de Ponderación de Interdependencia de Atributos: se utiliza cuando se desea identificar los factores con mayor incidencia sobre un determinado resultado de un problema.
- Proceso de Descubrimiento de Reglas de Pertenencia a Grupos: permite identificar las condiciones de pertenencia a cada una de las clases en una partición desconocida pero que se encuentra presente en la masa de información disponible sobre el dominio del problema.
- Proceso de Ponderación de Reglas de Comportamiento o de Pertenencia a Grupos: se utiliza cuando se requiere identificar las condiciones con mayor incidencia sobre la obtención de un determinado resultado en el dominio del problema, ya sea por la mayor medida en la que inciden sobre su comportamiento o las que mejor definen la pertenencia a un grupo.

A su vez, entre las tecnologías de sistemas inteligentes aplicadas a la explotación de información [García Martínez et al., 2003] se encuentran: los algoritmos de inducción o TDIDT, los mapas auto organizados de Kohonen o SOM (Self Organized Maps) y las redes bayesianas [Britos, 2008].

2.3. Método de Estimación para Proyectos de Explotación de Información (DMCoMo)

En [Marbán, 2003] se define un método analítico de estimación para proyectos de explotación de información el cual se denomina "Matemático Paramétrico de Estimación para Proyectos de Data Mining" (en inglés Data Mining COst MOdel, o DMCoMo). Este método, validado y desarrollado a través de una herramienta de software en [Pytel, 2011], permite estimar los meses/hombre que serán necesarios para desarrollar un proyecto de explotación de información desde su concepción hasta su puesta en marcha. Para realizar la estimación se definen seis categorías [Marbán, 2003] para vincular las características más importantes de los proyectos de explotación de información. Estas categorías son las siguientes: Datos, Modelos, Plataforma, Técnicas y Herramientas, Proyecto y Personal (Staff del Proyecto).

3. Definición del Problema

Al igual que en los proyectos de desarrollo de software tradicionales, los proyectos de Explotación de Información requieren de un proceso de planificación que permita estimar sus tiempos y medir el avance del producto en cada etapa de su desarrollo y calidad del mismo. Sin embargo, dada las diferencias que existen entre un proyecto clásico de construcción de software y un proyecto de explotación de información, las métricas usuales para realizar una estimación no se consideran completamente adecuadas ya que los parámetros a ser utilizados son de naturaleza diferentes y no se ajustan a sus características particulares, por ejemplo cantidad de fuentes de información, nivel de integración de los datos, el tipo de problema a ser resueltos, entre las más representativas de este tipo de proyectos. Como se mencionó en la sección anterior, el modelo de procesos para proyectos de Explotación de Información

propuesto en [Vanrell, 2011] plantea dos procesos principales: el proceso de administración de proyectos y el de desarrollo de proyectos de explotación de información. El proceso de administración de proyectos, se encarga de recolectar información necesaria para aumentar la calidad del proceso de desarrollo permitiendo realizar ajustes en el mismo y mantener un estándar en la realización de proyectos. Sin embargo, no plantea qué métricas utilizar para evaluar la calidad del proceso de desarrollo en este tipo de proyectos.

4. Solución Propuesta

En base a la categorización que en [Marbán, 2003] se realiza sobre el modelo de estimación DMCoMo, se plantea la utilización de aquellas categorías que sean aplicables al proceso de desarrollo de proyectos de Explotación de Información [Vanrell, 2011] como forma de clasificación de las métricas propuestas, focalizado este proceso en proyectos pequeños [Pytel, 2011] que son los que usualmente requieren las PyMEs. Estas métricas a su vez, se orientan a procesos de explotación de información que utilizan tecnologías de sistemas inteligentes.

4.1. Métricas de Datos

A partir del proceso de desarrollo de proyectos de Explotación de Información [Vanrell, 2011], se han considerado métricas de datos para las tareas que se indican en la tabla 1.

Tabla 1. Propuesta de Métricas de Datos

Subproceso: Entendimiento de los Datos
Tarea: Reunir los datos iniciales
Métricas Propuestas
▪ NT = Número total de fuentes de datos (tablas) necesarias para el proyecto. Se incluyen tablas internas y externas. ▪ NA (T)[1] = Número de atributos en la tabla T. ▪ NTA = Número total de atributos de las tablas. $NTA = \sum_{1}^{i} NA(T_i)$ ▪ NR (T)[1, 2]= Número de registros de la tabla T. Nota [1] – Métrica utilizada en la tarea: Explorar los datos y Limpiar los datos
Tarea: Explorar los datos
Métricas Propuestas
▪ NVN (T) = Número de valores nulos o faltantes en la tabla T. ▪ NANR (T) = Número de atributos nulos o faltantes del registro R en la tabla T. ▪ NCT (T) = Nivel de compleción de la tabla T. Mide el grado de compleción que tiene la tabla. $NCT(T) = 1 - \frac{NVN(T)}{NR(T)}$ ▪ DANR (T) = Densidad de atributos nulos o faltantes de un registro en la tabla T. Mide la proporción de atributos nulos o faltantes que tiene un registro R en la tabla T. $DANR(T) = \frac{NANR(T)}{NA(T)}$
Tarea: Verificar la calidad de los datos
Métricas Propuestas
▪ NAN (T) = Número de atributos a normalizar (transformados para su utilización) en la tabla T. ▪ NVE (A) = Número de valores erróneos del atributo A en la tabla. ▪ GIA (A) = Grado de integridad de un atributo A. $GIA(A) = 1 - \frac{NVE(A)}{NR(A)}$ donde NR es el número de registros que tiene el atributo A.

- NRN (T) = Número de registros con atributos nulos o faltantes en la tabla T.
- NRDE (T) = Número de registros con valores de atributos erróneos en la tabla T. Mide el nivel de precisión en los datos.
- NRVD (T) = Número de registros con valores duplicados en la tabla T.
- NRD (T) = Número de registros defectuosos en la tabla T. $NRD(T) = NRN(T) + NRDE(T) + NRVD(T)$

- GVRT (T) = Grado de validación de los registros de la tabla T. Mide el porcentaje de registros válidos obtenidos. Se recomienda que la validación obtenida sea al menos de 75%. $GVRT(T) = \frac{NR(T) - NRD(T)}{NR(T)} * 100$
- NVD = Nivel de volatilidad de los datos. Mide la frecuencia con que se cambian los datos en el tiempo [*0: no varían – 1: baja volatilidad – 2: volatilidad media – 3: alta volatilidad - 4: gran volatilidad*]

Subproceso: Preparación de los Datos

Tarea: Seleccionar los datos

Métricas Propuestas

- NAU (T) = Número de atributos útiles (significativos para el proyecto) y que no necesitan modificarse en la tabla T.
- NAM (T)[3] = Número de atributos útiles que se deben modificar (se incluye los atributos a normalizar) en la tabla T.

- Nota [3] – Métrica utilizada en la tarea: Limpiar los datos

Tarea Limpiar los datos

Métricas Propuestas

- NRE (T) = Número de registros eliminados en la tabla T.
- NAE (T) = Número de atributos a eliminar (no significativos para el proyecto) en la tabla T.
- GUT (T)[4] = Grado de utilidad de la tabla T. Mide el porcentaje de atributos útiles de la tabla T para el proyecto. A cada atributo se le asigna un peso según su estado de utilidad (*útiles = 0 / modificados = 0,5 / eliminados = 1*)

$$GUT(T) = \frac{NA(T) - (NAE(T) + 0{,}5 * NAM(T))}{NA(T)} * 100$$

Nota [4] – Métrica utilizada en la tarea: Integrar los datos

Tarea: Construir los datos

Métricas Propuestas

- NANI (TI) = Número de atributos nuevos a agregar en la integración de una única tabla TI.

Tarea: Integrar los datos

Métricas Propuestas

- NR (TI) = Número de registros de la tabla integrada.
- GUTAP = Grado de utilidad total de los atributos para el proyecto. Esta métrica mide el nivel de integración de todos los atributos disponibles en una única tabla. $GUTAP = \sum_{1}^{i} GUT(T_i)$

Si 0 <GUTAP < 40% los atributos no son usables.
Si 41% <GUTAP< 80% los atributos son aceptablemente usables.
Si 81 <GUTAP< 100% los atributos son muy usables.

4.2. Métricas de Modelos

En los proyectos de Explotación de Información es necesario evaluar la calidad de los modelos obtenidos de la manera más precisa que sea posible, para garantizar la aplicación de los mismos. Al no existir un modelo mejor que otro de manera general, para cada problema nuevo es necesario determinar con cuál se pueden obtener mejores resultados. A partir de los procesos de explotación de información definidos en [Britos, 2008] y según su tarea de descubrimiento, se pueden clasificar los modelos en: Descubrimiento de Grupos, Descubrimiento de Reglas y Descubrimiento de Dependencias Significativas.

A continuación se realiza una breve descripción de cada uno de estos modelos, las técnicas de sistemas inteligentes aplicables y se presentan los criterios escogidos para la evaluación de los modelos.

Descubrimiento de Grupos: Tiene por objetivo la separación de los datos en grupos (clusters) o clases basándose en la similitud de los valores de sus atributos. Todos los elementos del grupo deben tener características comunes pero a su vez entre los grupos los objetos deben ser diferentes [Britos, 2008]. Dentro de las

tecnologías inteligentes apropiadas para realizar agrupamiento están los mapas auto organizados de Kohonen (SOM – Self Organized Map, por sus siglas en inglés). Al construir un modelo de agrupamiento basado en mapas auto organizados, se define el número de grupos a priori, generando la necesidad de evaluar diferentes topologías para escoger de entre todas la mejor sub-óptima para la solución del problema. Estos mapas se basan en el aprendizaje no supervisado y competitivo. El factor de calidad del modelo generado está basado en el número de grupos que se definen al inicio, ya que al establecerse anticipadamente puede limitar la calidad de agrupamiento del algoritmo, y al ser una tarea de análisis exploratorio, no se sabe con precisión cuantos grupos pueden contener los datos.

Descubrimiento de Reglas: Es uno de los modelos más importantes de de la explotación de información. Se utiliza para encontrar las reglas de clasificación de un conjunto de elementos con base en los valores de sus atributos. El objetivo es lograr modelos de clasificación (expresados en reglas) que determinen correctamente la clase ante elementos no previstos anteriormente [Britos, 2008]. Los algoritmos mayormente utilizados para las tareas de clasificación son los algoritmos de inducción TDIDT (ID3, C4.5 y C5). Para evaluar un modelo de clasificación existen diversas métricas, sin embargo no es aconsejable emplear una sola de ellas ya que es común que una técnica de clasificación presente buenos resultados en una métrica y malos en otra. Por otra parte, al momento de aplicar las técnicas de clasificación se debe tener en cuenta cómo están distribuidos los elementos respecto a la clase o cluster. Puede ocurrir que al no estar balanceadas las clases los clasificadores estén sesgados a predecir un porcentaje más elevado de la clase más favorecida. Las métricas propuestas para evaluar un modelo de clasificación estarán basadas en la matriz de confusión que se obtiene cuando se prueba el clasificador en un conjunto de datos que no intervienen en el entrenamiento. Una matriz de confusión permite conocer la distribución del error a lo largo de las clases o clusters, cuando se prueba un clasificador en un conjunto de datos que no intervienen en el entrenamiento. Una matriz de confusión general tiene la siguiente estructura:

		Clase (cluster) predicha		
		Si	No	Totales
Clase (cluster) real	Si	Verdaderos Positivos (VP)	Falsos Negativos (FN)	Total Positivos Reales (TPR)
	No	Falsos Positivos (FP)	Verdaderos Negativos (VN)	Total Negativos Reales (TNR)

Los valores que se encuentran a lo largo de la diagonal principal de la matriz, representan las clasificaciones correctas y los que están a lo largo de la diagonal secundaria representan los errores (la confusión) entre las clases.

Descubrimiento de Dependencias Significativas: Consiste en encontrar modelos que describan dependencias o asociaciones significativas entre los datos. Las dependencias pueden ser usadas como valores de predicción de un dato, teniendo información de los otros datos. El análisis de dependencias tiene relación con la clasificación y la predicción, donde las dependencias están implícitamente usadas para la formulación de modelos predictivos [Britos, 2008]. Dentro de las técnicas de sistemas inteligentes apropiadas para realizar análisis de dependencias se encuentran las Redes Bayesianas. Si bien las redes bayesianas pueden utilizarse

dentro de los modelos de clasificación, hasta el momento no se han encontrado métricas significativas que establezcan un criterio adecuado de evaluación de dependencias ni de ponderación de atributos significativos.

Los modelos descriptos se corresponden con procesos de explotación de información basados en tecnologías de sistemas inteligentes unitarias [Britos, 2008]. En el caso del proceso de Descubrimiento de Reglas de Pertenencia a Grupos se necesita aplicar una combinación de los modelos de Descubrimiento de Grupos y Descubrimiento de Reglas; y en el caso del proceso de Ponderación de Reglas de Comportamiento o de Pertenencia a Grupos una combinación de los modelos de Descubrimiento de Grupos, Descubrimiento de Reglas y Descubrimiento de Dependencias Significativas, respectivamente. A partir del proceso de desarrollo de proyectos de Explotación de Información [Vanrell, 2011] y de la clasificación de los modelos establecida anteriormente, se han considerado métricas de modelos para las tareas que se indican en la tabla 2.

Tabla 2. Propuesta de Métricas de Modelos

Subproceso: Modelado
Tarea: Seleccionar técnica de modelado
Métricas Propuestas
▪ NM = Número de modelos a construir para el proyecto. ▪ NE (M) = Número de elementos (registros o casos) en el modelo M. ▪ NA (M) = Número de atributos en el modelo M. - PAN (M) = Porcentaje de atributos numéricos en el modelo M. - PANN (M) = Porcentaje de atributos no numéricos en el modelo M.
Tarea: Generar el diseño del test
Métricas Propuestas
▪ NEE (M) [5]= Número de elementos a utilizar para el entrenamiento del modelo M. ▪ NEP (M) [5] = Número de elementos a utilizar para las pruebas del modelo M. Nota [5] – Métrica utilizada en la tarea: Evaluar el modelo
Tarea: Construir el modelo
Métricas Propuestas
▪ NMDM (M) [6]= Número de modelos de explotación de información aplicados para construir el modelo M. Los modelos pueden ser: descubrimiento de grupos, descubrimiento de reglas y descubrimiento de dependencias significativas. Esta métrica tomará el valor 1, 2 ó 3, según la cantidad de modelos aplicados. Nota [6] – Métrica utilizada en la tarea: Evaluar los resultados
Tarea: Evaluar el modelo
Métricas Propuestas
o Modelo de Descubrimiento de Grupos ▪ NC (M) = Número de agrupamientos (cluster o clases) generados en el modelo M. ▪ NR (C) = Número de elementos agrupados en el cluster C. ▪ PR (C) = Porcentaje de elementos agrupados en el cluster C, respecto del total. ▪ NRP (C) = Número de elementos utilizados para las pruebas del cluster C. ▪ NEPE (M) = Número de elementos del conjunto de prueba del modelo M que fueron incorrectamente agrupados. ▪ TEMA (M) = Tasa de error del modelo de agrupamiento o clustering. Mide el porcentaje de elementos de prueba que fueron mal agrupados. $TEMA(M) = \frac{NEPE(M)}{NEE(M) + NEP(M)} * 100$ ▪ EMA (M) = Nivel de exactitud del modelo de agrupamiento. $EMA(M) = 1 - \frac{TEMA(M)}{100}$ o Modelo de Descubrimiento de Reglas ▪ VP (C) = Número de elementos positivos del cluster C clasificados correctamente. ▪ VN (C) = Número de elementos negativos del cluster C clasificados correctamente. ▪ FP (C) = Número de elementos negativos del cluster C clasificados incorrectamente. ▪ FN (C) = Número de elementos positivos del cluster C clasificados incorrectamente. ▪ IE (C) = Índice de pertenencia de un elemento al cluster C. Mide la probabilidad de que un elemento pertenezca a un determinado cluster C. $IE(C) = \frac{VP(C)}{NR(C)}$

- TPR (C) = Número total de elementos positivos reales del cluster C. $TPR(C) = VP(C) + FN(C)$
- TNR (C) = Número total de elementos negativos reales del cluster C. $TNR(C) = FP(C) + VN(C)$
- Acierto: es la proporción del número total de casos predichos que son correctos. Mide el nivel de certeza del modelo.

$$\text{Acierto} = \frac{VP + VN}{TPR + TNR}$$

- SMC (M) = Sensibilidad del modelo de clasificación. Mide la proporción de casos positivos que fueron identificados correctamente, es decir la probabilidad que un elemento de una clase o cluster sea clasificado correctamente en esa misma.

$$SMC(M) = \frac{VP}{TPR} = \text{TasadeVerdaderosPositivos}$$

- ESMC (C) = Especificidad del modelo de clasificación. Mide la proporción de casos negativos que fueron identificados correctamente, es decir, la probabilidad que un elemento de una clase o cluster sea clasificado correctamente en la misma.

$$ESMC(C) = \frac{VN}{TNR} = \text{TasadeVerdaderosNegativos}$$

- PMC (M) = Precisión del modelo de clasificación. Mide la proporción de los casos predichos positivos que son correctos, es decir, la probabilidad de que una predicción efectivamente corresponda con su valor real. $PMC(M) = \frac{VP}{VP + FP}$
- EXMC (M): Exactitud del modelo de clasificación. Mide la proporción de casos correctamente predichos.

$$EXMC(M) = \frac{VP + VN}{TPR + TNR}$$

- Error ó Confusión: es la proporción de casos incorrectamente predichos. $\text{Error} = \frac{FP + FN}{TPR + TNR}$
- Tasa de Falsos Positivos y Negativos: es la proporción de casos incorrectamente predichos.

$$\text{TasadeFalsosPositivos} = \frac{FP}{TNR} \qquad \text{TasadeFalsosNegativos} = \frac{FN}{TPR}$$

- Medida F: es una medida estadística que combina las métricas de Precisión y Sensitividad para evaluar de forma más realista la certeza del modelo. $\text{MedidaF} = \frac{2}{\frac{1}{PMC(M)} + \frac{1}{SMC(M)}}$
- Coeficiente de Kappa: mide la precisión del modelo para predecir la clase o cluster verdadero. $k = \frac{P(\text{Acierto}) - P(\text{Error})}{1 - P(\text{Error})}$

  donde P (Acierto) es el porcentaje de aciertos y P (Error) es el porcentaje de casos incorrectamente clasificados.

o Modelo de Descubrimiento de Dependencias Significativas

En esta instancia de la investigación no se han encontrado métricas representativas para este modelo.

4.3. Métricas de Proyecto

Si bien en [Marbán, 2003] se definen las categorías que involucran las características más importantes de los proyectos de Explotación de Información, dentro de la cual se encuentra la categoría Proyecto, no se hace ninguna mención en ésta sobre los criterios de evaluación de los resultados obtenidos del proceso de minería de datos, para lograr un resultado exitoso del proyecto. Con lo cual, las métricas que se proponen para evaluar los resultados, se incluirán dentro de esta categorización. En la tarea de evaluación de los modelos, se estableció la métrica de exactitud como uno de los factores para evaluar la calidad de los modelos construidos. La exactitud de un modelo está relacionada con el grado de confiabilidad que tienen los resultados obtenidos frente al objetivo de minería de datos que se persigue y, en consecuencia, con el éxito del proceso de desarrollo del proyecto de explotación de información. Para definir la métrica que determina el éxito de un proyecto de explotación de información, se considerará el resultado de la métrica de exactitud para cada modelo M de minería de datos construido. A cada modelo M, se le asignará un peso según su nivel de exactitud, como se indica en la tabla 3:

Tabla 3. Pesos asociados a niveles de exactitud de modelos

NIVEL DE EXACTITUD	RANGO DE LA MÉTRICA	PESO
Muy Bajo	0 – 20%	0
Bajo	21 – 49%	0.25
Nominal	50 – 70%	0.5
Alto	71 – 90%	0.75
Muy Alto	> 90%	1

A partir del proceso de desarrollo de proyectos de Explotación de Información [Vanrell, 2011], se han considerado métricas de proyecto para las tareas que se indican en la tabla 4.

Tabla 4. Propuesta de Métricas de Proyecto

Subproceso: Evaluación
Tarea: Evaluar los resultados
Métricas Propuestas
▪ NM = Número de modelos construidos para el proyecto. ▪ NEM (M) = Nivel del exactitud del modelo M. Mide el nivel de exactitud de cada modelo construido para el proyecto de explotación de información. $NEM(M) = \frac{\sum Exactitud\_Modelos(M)}{NMDM(M)} * 100$ donde Exactitud_Modelos son los niveles de exactitud obtenidos para cada clasificación de modelo de explotación de información (descubrimiento de grupos, descubrimiento de reglas y descubrimiento de dependencias significativas). A cada modelo M, se le asigna un peso según la tabla 4. ▪ NMMA = Número de modelos clasificados como Muy Alto. ▪ NMA = Número de modelos clasificados como Alto. ▪ NMN = Número de modelos clasificados como Nominal. ▪ NMB = Número de modelos clasificados como Bajo. ▪ Éxito = Éxito del proceso de desarrollo. Mide el nivel de los resultados de minería de datos obtenidos respecto a los criterios de éxito del proyecto. Cuanto mayor sea el este valor, implicará una mayor aceptación por parte del usuario del proyecto de explotación de información. $Éxito = \frac{NMMA + (NMA * 0.75) + (NMN * 0.5) + (NMB * 0.25)}{NM} * 100$
Tarea: Revisar el proceso
Métricas Propuestas
▪ Si 0 <Éxito < 50% el proceso no responde a los criterios de éxito del problema de negocio. ▪ Si 51% <Éxito< 80% el proceso debe ser revisado y ajustado. ▪ Si 81 <Éxito< 100 el proceso cumple con los criterios de éxito del problema de negocio.
Subproceso: Entrega
Tarea: Producir un reporte final
Métricas Propuestas
▪ GDE = Grado de documentación a entregar. Mide el esfuerzo necesario para producir la documentación durante el proyecto. Los valores pueden ser: - Bajo: Sólo para los documentos implantados - Nominal: Sólo para modelos generados - Alto: Todos los modelos (generados e implantados) y subprocesos del proceso de desarrollo. - Muy Alto – Todos los modelos (generados e implantados) y procesos del proyecto de explotación de información (administración y desarrollo).

5. Conclusiones

Se definió una propuesta de métricas aplicable en proyectos de explotación de información para PyMEs, que permite evaluar su avance y calidad durante el proceso de desarrollo. Del trabajo realizado, surge que algunas tareas definidas en el proceso

de desarrollo [Vanrell, 2011] no permiten obtener métricas significativas que deban ser consideradas.

Se prevé tener una primer validación de las metricas propuestas en el marco de los proyectos de explotación de información que desarrollan los alumnos en la Asignatura "Tecnologías para Explotación de Información" en la Carrera de Licenciatura en Sistemas de la Universidad Nacional de Lanús y en la Carrera de Ingeniería en Sistemas de Información de la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional.

Como futura línea de trabajo se estima continuar con la definición de métricas aplicables a la ponderación de atributos en el Modelo de Dependencias para Explotación de Información, y la validación de la solución propuesta en experiencias de trabajo tanto académicas como reales.

6. Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el Proyecto de Investigación 33A167 de la Secretaria de Ciencia y Técnica de la Universidad Nacional de Lanús (Argentina); y por el Programa de Incentivos a la Investigación del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza (Argentina).

7. Referencias

Britos, P. 2008. Procesos de Explotación de Información Basados en Sistemas Inteligentes. Tesis Doctoral en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata.

García Martínez, R., Servente, M. y Pasquini, D., 2003. Sistemas Inteligentes. Editorial Nueva Librería. Buenos Aires.

Hauge, O., Britos, P., García-Martínez, 2006. Conceptualization Maturity Metrics for Expert Systems. IFIP International Federation for Information Processing.

Marbán, O. 2003. Modelo Matemático Paramétrico de Estimación para Proyectos de Data Mining (DMCoMo). Tesis Doctoral. Facultad de Informática. Universidad Politécnica de Madrid.

Pollo Cataneo, M.F, 2007. Sistemas Expertos. Conceptualización y Métrica de Madurez. Trabajo Final de Especialidad en Ingeniería de Sistemas Expertos. ITBA Instituto Tecnológico de Buenos Aires.

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

Vanrell, J, 2011. Un Modelo de Procesos para Proyectos de Explotación de Información. Tesis de Magister en Ingeniería de Sistemas de Información. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional.