Propuesta de Métricas para Proyectos de Explotación de Información

Diego Martín Basso [1, 2, 3]
1.Maestría en Ingeniería de Sistemas de Información. Universidad Tecnológica Nacional, FRBA
Buenos Aires, Argentina
2. Grupo de Investigación en Sistemas de Información. Universidad Nacional de Lanús
Remedios de Escalada, Argentina
3. Departamento de Ingeniería e Investigaciones Tecnológicas. Universidad Nacional de La Matanza
San Justo - Argentina
diebasso@yahoo.com.ar

Resumen **- Los Proyectos de Explotación de Información requieren de un proceso de planificación para estimar el esfuerzo, el tiempo y medir diferentes aspectos del producto para garantizar la calidad del mismo. Los procesos de desarrollo tradicionales y las métricas usuales de la Ingeniería de Software y la Ingeniería del Conocimiento no son adecuados para estos proyectos, ya que las etapas de desarrollo y los parámetros utilizados son de naturaleza y características diferentes. En ese contexto, se ha definido un Modelo de Proceso de Desarrollo para Proyectos Explotación de Información. No obstante, existe la necesidad de abordar métricas específicas aplicables a este proceso. En esta investigación se propone un conjunto de métricas aplicables al desarrollo de un proyecto de Explotación de Información para PyMEs, centrado en el Modelo de Proceso de Desarrollo mencionado.**

Palabras Clave **- Inteligencia de Negocios, Explotación de Información, Ingeniería de Proyectos de Explotación de Información, Métricas, Minería de Datos.**

I. Introducción

En este apartado se presenta el contexto de la investigación (sección *A*), esbozando uno de los problemas abiertos identificados (sección *B*), se mencionan los objetivos (sección *C*), en particular el objetivo general (sub-sección *C.1*), los objetivos específicos (sub-sección *C.2*) y el alcance de los mismos (sub-sección *C.3*), y se describe la metodología empleada para el desarrollo del trabajo (sección *D*). El apartado finaliza con la descripción de la estructura general de la investigación (sección *E*) y el resumen de las producciones científicas vinculadas a la misma (sección *F*).

A. Contexto de la Investigación

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

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

En el ámbito de la Ingeniería en Conocimiento, especialmente en el desarrollo de sistemas expertos o sistemas basados en conocimientos, un aspecto importante que debe ser medido es la conceptualización para poder estimar actividades futuras y obtener información del estado de madurez del conocimiento sobre el dominio y sus particularidades [27]. Estas métricas de madurez de conceptualización para Sistemas Expertos definidas en [27] y aplicadas en [61] [62] brindan además información sobre la complejidad del dominio.

Los Proyectos de Explotación de Información también requieren de un proceso de planificación que permita estimar sus tiempos y medir el avance del producto en cada etapa de su desarrollo y calidad del mismo. Sin embargo, como consecuencia de las diferencias que existen entre un proyecto clásico de construcción de software y un proyecto de explotación de información [84], las métricas de software usuales no se consideran adecuadas ya que los parámetros utilizados en los proyectos de explotación de información son de naturaleza diferentes [48] [49] y no se ajustan a sus características específicas.

B. Problemas Abiertos

En virtud a la comprobación que los procesos de desarrollo tradicionales de la Ingeniería de Software no son aptos para desarrollar Proyectos de Explotación de Información [85] y ante falta de técnicas asociadas a la ejecución de cada una de las fases de las metodologías de explotación de información vigentes [24], se ha desarrollado un Modelo de Proceso de Desarrollo aplicable a Proyectos de Explotación de Información [85]. En ese sentido, se ha señalado la necesidad de disponer de métricas asociadas a este proceso [24] que permitan evaluar la calidad del proceso, el producto entregable y realizar un correcto seguimiento del proyecto.

C. Objetivos de la Investigación

Se dividen los objetivos de esta investigación en un objetivo general a alcanzar (sub-sección *1*), un conjunto de objetivos específicos (sub-sección *2*) que definen los pasos a

seguir para lograr el objetivo general y el alcance previsto con estos objetivos (sub-sección *3*).

1. Objetivo General

El objetivo de esta investigación es definir una propuesta de métricas aplicables al proceso de desarrollo de Proyectos de Explotación de Información para PyMEs (Pequeñas y Medianas Empresas), que permitan medir diversos atributos del proyecto.

2. Objetivos Específicos

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

- Establecer las categorías aplicables al proceso de desarrollo de Proyectos de Explotación de Información, como forma de clasificación de las métricas a proponer.
- Proponer un conjunto de métricas significativas para Proyectos de Explotación de Información siguiendo los lineamientos, subprocesos y tareas enunciadas en el Modelo de Proceso Desarrollo definido en [84].
- Analizar y estudiar el comportamiento de las métricas propuestas utilizando un método de validación empírico por simulación.

3. Alcance de los Objetivos Específicos

Se detalla a continuación el alcance esperado con los objetivos específicos en el ámbito de la Ingeniería de Proyectos de Explotación de Información:

- Las métricas propuestas deben considerar la aplicación de los procesos de explotación de información definidos por [5], los cuales utilizan tecnologías de sistemas inteligentes [23].
- La propuesta de métricas debe ser escalable, a fin de permitir la incorporación de nuevas métricas.

D. Metodología Empleada

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

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

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

En una etapa posterior, y a partir de los métodos de estimación de esfuerzo para proyectos de explotación de información y los procesos existentes para realizar la tarea de explotación, se establecen las categorías de métricas para el Modelo de Proceso de Desarrollo.

En el siguiente paso y utilizando como marco de referencia este modelo de proceso de desarrollo y sus tareas asociadas, se proponen las métricas para este trabajo de investigación.

A continuación se analiza y estudia el comportamiento de las métricas definidas, mediante un método de validación empírico por simulación Monte Carlo.

Finalmente se elabora el informe final con las conclusiones obtenidas y se establecen futuras líneas de trabajo para continuar con esta investigación.

E. Estructura General de la Investigación

La investigación se estructura en seis apartados: Introducción, Estado de la Cuestión, Descripción del Problema, Solución Propuesta, Estudio del Comportamiento de las Métricas y Conclusiones, los cuales se describen a continuación:

En el *apartado I - Introducción* se plantea el contexto que da soporte a este trabajo de investigación, se establecen los objetivos, en particular el objetivo general, los objetivos específicos y el alcance de los mismos y se describe la metodología empleada para el desarrollo del trabajo. El apartado finaliza describiendo la estructura general de la investigación y resumiendo las producciones científicas vinculadas con la misma.

En el *apartado II - Estado de la Cuestión* se plantea la importancia que tienen las métricas en el desarrollo de proyectos y para el aseguramiento de la calidad, se clasifican las métricas en base a los aspectos que pueden medirse en el software y se identifican diversas métricas utilizadas en la Ingeniería de Software y la Ingeniería del Conocimiento, especialmente en el desarrollo de los Sistemas Expertos. Posteriormente, se describe el contexto de la Ingeniería de Proyectos de Explotación de Información, identificando las diferencias de esta disciplina con la Ingeniería del Software. Asimismo, se mencionan las particularidades de los métodos propuestos para estimación de esfuerzo en proyectos de explotación de información y los procesos definidos para realizar esta explotación. Finalmente, se menciona el modelo de procesos que cubriría el desarrollo de proyectos de explotación, con aplicación a empresas PyMEs, y se plantea la necesidad de definir métricas específicas para estos proyectos.

En el *apartado III - Descripción del Problema* se desarrolla la problemática que intenta solucionar esta investigación junto con la justificación de las decisiones tomadas para llevar a cabo su resolución, finalizando con las preguntas de investigación que se intentaron responder mediante este trabajo de investigación.

En el *apartado IV - Solución* se desarrolla la solución propuesta para la problemática mencionada en el apartado anterior, realizando una descripción de las consideraciones establecidas durante esta propuesta, para finalmente proponer el conjunto de métricas que dan cumplimiento a los objetivos general y específico de la investigación.

En el *apartado V – Estudio del Comportamiento de las Métricas* se analiza y estudia el comportamiento de las métricas propuestas como solución, utilizando un método empírico de simulación por Monte Carlo. Para las métricas sometidas a estudio, se obtienen las conclusiones generales de su comportamiento dentro de ámbito de los proyectos de explotación de información.

En el *apartado VI – Conclusiones* se describen las conclusiones obtenidas a partir del desarrollo de esta investigación y se mencionan las aportaciones obtenidas al cuerpo de conocimiento de la Ingeniería de Proyectos de Explotación de Información. Asimismo, se da respuesta a los interrogantes de investigación planteados, finalizando el

apartado con las futuras líneas de investigación a desarrollar a partir de esta investigación.

Se finaliza el trabajo enunciando todas las referencias bibliográficas utilizadas para esta investigación.

II. Estado de la Cuestión

Este apartado describe cómo es el contexto de trabajo en cual se inserta esta investigación, permitiendo al lector adquirir los conocimientos necesarios para comprender la problemática de la misma, como así también el por qué de la solución planteada. En el mismo se indica la importancia que tienen las métricas en el desarrollo de proyectos (sección *A*) y su relación con la calidad, se identifican algunas métricas existentes (sección *B*) en el ámbito de la Ingeniería de Software y la Ingeniería del Conocimiento, en especial en el desarrollo de los Sistemas Expertos, finalizando el apartado con una contextualización de la Ingeniería de Proyectos de Explotación de Información y la problemática de no contar con métricas específicas para estos proyectos (sección *C*).

A. Importancia de las Métricas en el Desarrollo de Proyectos

En la actualidad, existe un importante interés por parte de las empresas que desarrollan software por lograr que los productos software cumplan con ciertos indicadores de calidad en todas las etapas del desarrollo [60]. Con independencia del tipo de producto que se desarrolle en un proyecto, la calidad es fundamental para lograr la satisfacción de las necesidades y expectativas del cliente. En [63] se menciona que el aseguramiento de la calidad del software (SQA) es una "actividad de protección" que se aplica a lo largo de todo el proceso de Ingeniería de Software, en la que se incluyen mecanismos para medir el producto y el proyecto, entre otros.

Dentro de los estándares que proporcionan modelos para la evaluación de la calidad del software, se encuentran las normas ISO 9000 (especialmente ISO 9001 e ISO 9003-2) [72], CMM (Capability Maturity Model) [75] y CMMI (Capability Maturity Model Integration) [79]. Estos modelos incorporan técnicas y procesos para el aseguramiento de la calidad que se corresponden con la medición del software. Por otra parte, algunos autores [51] [63] y estándares internacionales [30] [34], han tratado de determinar y categorizar las características que deben cumplir todo producto software para ser considerado de calidad, y a partir de éstas proporcionar la terminología para especificar, medir y evaluar la calidad del mismo.

La medición de un software es tan importante en cualquier proceso de ingeniería como su misma construcción [13], ya que permite tener una visión del proyecto, de la evaluación del producto y de su nivel de aceptación, logrando un mejoramiento continuo del software y permitiendo cuantificar y gestionar, de forma más efectiva, cada una de las variables a las que se necesite hacer seguimiento. En este sentido, la medición del software debe satisfacer tres objetivos fundamentales [15]: (1) ayudar a entender qué ocurre durante el desarrollo y el mantenimiento, (2) permitir controlar qué es lo que ocurre en los proyectos y (3) poder mejorar los procesos y productos. Mejorar la calidad de los resultados de un proyecto de software o la eficiencia de sus procesos es difícil, si no se recolectan métricas.

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

Los proyectos de Explotación de Información también deben considerar la aplicación de una metodología de desarrollo [24] que incluya entre sus actividades el registro de métricas, que permitan medir y controlar el avance del proyecto y evaluar su calidad. Asimismo, se destaca la necesidad de realizar estimaciones de esfuerzo al comienzo de estos proyectos y compararlos con los valores reales al finalizar el mismo.

Dentro del cuerpo de conocimiento de la Ingeniería de Proyectos de Explotación de Información, se han propuesto y desarrollado distintas herramientas [24], entre las que se mencionan: un modelo de procesos, un proceso de educción de requisitos, un método de estimación, una metodología de selección de herramientas, un proceso de transformación de datos y una serie de procesos basados en técnicas de minería de datos. Estas herramientas además, han sido utilizadas en proyectos para pequeños y medianos emprendimientos. Asimismo, en el trabajo de [24] se ha señalado la necesidad de plantear métricas significativas asociadas al proceso de desarrollo de un proyecto de Explotación de Información, que permitan suministrar información relevante a tiempo y establecer objetivos de mejora en los procesos y productos, con el fin de garantizar la calidad de estos proyectos.

Se presenta continuación la definición del concepto de métrica (sub-sección *1*) y su relación con las medidas, indicadores y medición, y se identifican los atributos que se pueden medir en un proyecto de desarrollo de software (sub-sección *2*).

1. Definición de Métrica

Una métrica se puede definir básicamente como la medición numérica de un atributo ante la necesidad de tener información cuantitativa del mismo para la toma de decisiones. En los proyectos de desarrollo de software, a menudo se suele hablar de métricas y de medidas, indistintamente. Sin embargo, existen diferencias entre estos términos. Se proporcionan a continuación algunas definiciones que permiten entender el concepto de métrica.

- *Medida*: valor asignado a un atributo de una entidad mediante una medición. Es una medida que proporciona una indicación cuantitativa de extensión, cantidad, dimensiones, capacidad y tamaño de algunos atributos de un proceso o producto [63].
- *Métrica*: medida cuantitativa del grado en que un sistema, componente o proceso posee un determinado atributo [29] [63].
- *Indicador*: es una métrica o una combinación de métricas que proporcionan una visión profunda del proceso del software, del proyecto de software, o del producto en sí [63].
- *Medición*: es el proceso por el cual los valores son asignados a atributos o entidades en el mundo real tal como son descritos de acuerdo a reglas claramente definidas [16]. Dicho de otro modo, es el proceso por el cual se obtiene una medida [61].

Estas definiciones permiten afirmar que a partir de los valores de las medidas, es posible reunir métricas que proporcionen información mediante indicadores, para poder controlar la eficacia del proceso, del proyecto o del producto software [61].

2. Atributos Medibles en el Software

Cualquier cosa que se quiera medir o predecir en un software representa un atributo (propiedad) de cualquier entidad de un producto, proceso o recurso asociado a éste. Cada entidad de software tiene varios atributos internos y externos que pueden ser medidos [15].

Los atributos internos de un producto, proceso o recurso son aquellos que se pueden medir directamente en términos del producto, proceso o recurso del mismo [16], por ejemplo: el tamaño del software, el esfuerzo para desarrollar un módulo del software, el tiempo transcurrido en la ejecución de cualquier módulo de software, entre otros. Los atributos externos de un producto, proceso o recurso son aquellos que solamente pueden ser medidos con respecto a cómo el producto, proceso o recurso se relacionan con su entorno [63], por ejemplo: el costo de eficacia de los procesos, productividad del grupo de desarrollo, complejidad del proyecto, la usabilidad, fiabilidad, o portabilidad de un sistema, entre otros. Los atributos externos son los más difíciles de medir, porque estos no pueden ser medidos directamente [16].

Los valores de los atributos se obtienen tras realizar mediciones sobre el software. Las mediciones dan como resultado una serie de métricas, que según la norma ISO/IEC 9126 [31] [32] [33] se pueden clasificar en tres categorías, según sea su naturaleza:

- *Métricas básicas*: son métricas que se obtienen directamente del análisis del código o la ejecución del software. No involucra ningún otro atributo ni depende de otras métricas. En [63] estas métricas se denominan directas. Entre las métricas básicas se tiene la cantidad de líneas de código del programa o de cada módulo, la cantidad de horas de desarrollo, la cantidad de fuentes de datos o tablas a utilizar, la cantidad de atributos y registros de una tabla, entre otras.
- *Métricas de agregación*: son métricas compuestas a partir de un conjunto definido de métricas básicas (o directas), generalmente mediante una suma ponderada.
- *Métricas derivadas*: son métricas compuestas por una función de cálculo matemático, que utiliza como variables de entrada el valor de otras métricas. En [63] estas métricas se denominan indirectas. Entre las métricas derivadas se tiene la cantidad de líneas de código producidas por hora y por persona, el porcentaje de completitud del proyecto, el tamaño promedio de los módulos del software, el tiempo promedio que una persona dedica a corregir los defectos de un módulo, entre otras.

B. Métricas Existentes

Para medir el desarrollo de un proyecto es necesario saber qué entidades son medidas y tener una idea de los atributos de la entidad. Para ello, se debe identificar el atributo a medir y su significado de medición [56]. Por otra parte, diversos autores han propuesto distintos tipos de métricas de acuerdo a la relevancia de lo que se esté midiendo.

Para poder llegar a un buen resultado en un proyecto de software se debe comprender el dominio del problema donde se va a trabajar, los recursos necesarios, las actividades y tareas a llevar a cabo, el esfuerzo y tiempo que se va a insumir, el plan de acción y los riesgos que se van a correr [61]. En este sentido, la naturaleza, el tamaño del proyecto y el entorno en el que se desarrolla son factores determinantes y afectan en gran medida a la estimación que se realice. Dentro del campo de la Ingeniería de Software (sub-sección *1*) e Ingeniería del Conocimiento (sub-sección *2*) existen diferentes tipos de métricas, que pueden ser clasificadas según los aspectos que miden.

1. Métricas en Ingeniería de Software

Las métricas de software proporcionan información relevante a tiempo que contribuye a gestionar de forma más efectiva un proyecto, y mejorar la calidad de los procesos y productos de software [63]. Además, al conocerse el estado actual del desarrollo de un proyecto, pueden establecerse objetivos de mejora [36]. Por otra parte, el uso de métricas no sólo permite entender, monitorizar, controlar, predecir y probar el desarrollo de software y los proyectos de mantenimiento [4] sino que también pueden ser utilizadas para tomar mejores decisiones [59].

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

Las métricas de software contemplan varias clasificaciones que apuntan a diferentes aspectos del proceso y del producto de software [63]. De acuerdo al contexto o dominio de aplicación y de las características o atributos del software, las métricas de software se pueden tipificar en: métricas del producto, del proceso y del proyecto. A su vez, algunas de estas métricas pueden pertenecer a más de una clasificación.

a) Métricas del Producto: son métricas que evalúan la calidad de los productos entregables, permitiendo tener un conocimiento detallado del diseño y la construcción del producto software. En estas métricas se tienen en cuenta atributos como: tamaño, calidad, complejidad, esfuerzo, volatilidad, entre otros.

b) Métricas del Proceso: son métricas aplicadas a fines estratégicos y propician indicadores que conducen a avances en el proceso y ambiente de desarrollo del software, a partir de información histórica de procesos similares. Se utilizan para evaluar si la eficiencia de un proceso ha mejorado en el largo plazo. Se recopilan de todos los proyectos y durante un largo período de tiempo. Dentro de estas métricas se incluye atributos como la experiencia del grupo, el costo del desarrollo y mantenimiento, el esfuerzo y tiempo dedicado a las pruebas, tiempo de desarrollo (total y por proceso, subproceso), tipo y cantidad de fallas, número de cambio con modelos previos, costo de aseguramiento de la calidad, cantidad de personas por día, por mes, intensidad del trabajo, interrupciones, entre otros [78].

c) Métricas del Proyecto: son métricas de tipo tácticas y describen las características propias del proyecto y de su ejecución. Estas métricas reducen la planificación del desarrollo ya que permite realizar los ajustes necesarios para evitar retrasos o riesgos potenciales, minimizar los defectos y por lo tanto la cantidad de trabajo que debe rehacerse, ocasionando en consecuencia una reducción del costo global

del proyecto [52]. A su vez, permiten evaluar la calidad de los productos obtenidos en cada etapa del desarrollo [52]. Estas métricas tienen en cuenta atributos como duración real del proyecto, esfuerzo real [persona-mes] por proceso, subproceso y por proyecto, progreso del proyecto, tamaño del proyecto, costo total invertido, entre otros.

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

2. Métricas en Ingeniería del Conocimiento

La planificación juega un papel esencial en la gestión de un proyecto de software, en la que se debe estimar el esfuerzo humano, costo y tiempo. Para esto se tienen métricas de software, que permiten obtener información y así generar conocimiento de la evolución y alcance del proyecto. En el desarrollo de Sistemas Expertos o Sistemas Basados en Conocimiento, la planificación presenta particularidades que la hacen altamente compleja [62]. Como se mencionó, proceso y producto son elementos protagonistas de las técnicas de medición. En este sentido, en el trabajo de [27] se han propuesto un conjunto de métricas de madurez aplicables en la fase de conceptualización, que examinan el dominio del problema en el contexto de desarrollo de un Sistema Experto [17]. Estas métricas, aplicadas en [61] [62], brindan además información sobre la madurez de la base de conocimientos y la complejidad del dominio. Se presentan a continuación las métricas existentes definidas en [27], las cuales se basan en *Reglas*, *Conceptos*, *Atributos* y *Niveles de Descomposición*.

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

C. Contexto de los Proyectos de Explotación de Información

La Explotación de Información (en inglés Information Mining, IM) es la sub-disciplina de los sistemas de información vinculada a la Inteligencia de Negocio [55] que aporta las herramientas de análisis y síntesis para extraer conocimiento, que se encuentra de manera implícita en los datos disponibles de diferentes fuentes de información [77]. Dicho de otra manera, las herramientas que permiten transformar la información en conocimiento [1] [11] [24] [53] [80]. En [43] se define a la Explotación de Información como el proceso de descubrir nuevas correlaciones, patrones y tendencias significativas utilizando grandes cantidades de datos almacenados en repositorios, usando tecnologías de reconocimiento de patrones, así como técnicas matemáticas y de estadística. En este contexto, la Ingeniería de Proyectos de Explotación de Información estudia los procesos de extracción de conocimiento no trivial [50], el cual es previamente desconocido y puede resultar útil para algún proceso [81].

Un proceso de Explotación de Información se define como un conjunto de tareas relacionadas lógicamente [12] [25], el cual engloba un conjunto de técnicas de minería de datos (en inglés Data Mining, DM) que pueden ser elegidas para realizarlas y así lograr extraer de conocimiento procesable, implícito en el almacén de datos (en inglés Data Warehouse, DW) de la organización. Las bases de estas técnicas se encuentran en el análisis estadístico y en los sistemas inteligentes. De esta manera, se aborda la solución a problemas de predicción, clasificación y segmentación [83].

Un proyecto de Explotación de Información involucra, en general las siguientes fases [47]: comprensión del negocio y del problema que se quiere resolver, determinación, obtención y limpieza de los datos necesarios, creación de modelos matemáticos, ejecución, validación de los algoritmos, comunicación de los resultados obtenidos, e integración de los mismos, si procede, con los resultados en un sistema transaccional o similar. La relación entre todas estas fases tiene una complejidad que se traduce en una jerarquía de subfases.

Por otra parte, y a partir de la experiencia adquirida en proyectos de Explotación de Información, se han desarrollado diferentes metodologías de desarrollo que permiten gestionar esta complejidad de una manera uniforme, siendo CRISP-DM [9], SEMMA [76] y P^3TQ [64] las metodologías probadas por la comunidad científica [5] [21].

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

Sin embargo, las fases habituales de un proyecto clásico de desarrollo de software (análisis, diseño, construcción, integración y prueba) no encuadran con las etapas propias de un proyecto de Explotación de Información [84].

Esto significa que las herramientas clásicas de la Ingeniería de Software tales como la ingeniería de requerimientos, los modelos de procesos, los ciclos de vida y los mapas de actividades no sean aplicables para los proyectos de Explotación de Información [24].

Por otra parte, un proyecto de Explotación de Información considera entre sus características más representativas la cantidad de fuentes de información a utilizar, el nivel de integración y calidad que presentan los datos, el tipo de proble ma de explotación de información a ser resuelto, los modelos necesarios a construir para el proyecto y la utilidad e interés de los patrones de conocimiento descubierto, entre otras.

TABLA I. MÉTRICAS EXISTENTES EN INGENIERÍA DE SOFTWARE

Aspectos del Software	Atributo	Métricas existentes
Producto	Tamaño	- Líneas de Código (medidas en miles - KLDC) - Puntos de Función (PF) - Páginas de Documentación
	Complejidad	- Complejidad ciclomática - Nivel de acoplamiento de los módulos -Nivel de modularidad (cohesión de módulos)
	Calidad	- Cantidad de defectos por KLDC - Cantidad de errores encontrados por KLDC - Cantidad de defectos/errores que encuentran los usuarios después de la entrega - Tipo y origen de los defectos (requerimientos, análisis y diseño, construcción, integración y pruebas)
	Mantenimiento	- Cantidad de componentes - Volatilidad de los componentes - Complejidad de los componentes - Cantidad de requerimientos nuevos, de cambios o mejoras - Cantidad de requerimientos de corrección de defectos - Tiempo promedio de corrección de errores ó defectos - Tiempo promedio de cambios - Porcentaje del código corregido
	Confiabilidad	- Tiempo transcurrido entre fallas - Tiempo esperado entre fallas - Tiempo requerido para corregir una falla - Nivel de severidad de la falla
	Usabilidad	- Facilidad de aprendizaje de uso - Errores cometidos por los usuarios con el uso - Tiempo requerido para realizar las tareas
	Rendimiento	- Tiempos de respuesta (acuerdos SLA) - Utilización de recursos (Troughput o Thruput) – cantidad de transacciones que pueden ejecutarse concurrentemente con un tiempo de respuesta razonable. - Tiempo de recuperación
Proyecto	Esfuerzo	- Cantidad de horas trabajadas - Cantidad de personas que trabajan en el proyecto - Tiempo transcurrido - Distribución del esfuerzo por fase
	Costo	- Costo del Desarrollo - Costo del Soporte - Costo de hs/persona
	Productividad	- Cantidad de software desarrollado por unidad de tiempo de trabajo - Tamaño/Esfuerzo - Ritmo de entrega del software por unidad de tiempo transcurrido.
	Seguimiento	- Cronograma real vs Cronograma estimado - Porcentaje de tareas completadas - Porcentaje de requerimientos implementados por unidad de tiempo - Porcentaje de tiempo total dedicado a las pruebas - Porcentaje de error en la estimación del tiempo - Costo sobre el valor agregado
	Estabilidad	- Origen de los cambios en los requerimientos - Cambios de los requerimientos en el desarrollo - Cambios en los requerimientos en producción
Proceso	Esfuerzo	- Distribución del esfuerzo por fase del proceso - Cantidad de personas requeridas - Esfuerzo requerido para corregir un defecto - Esfuerzo requerido para mejorar un defecto
	Reusabilidad	- Cantidad de componentes reutilizados - Grado de reusabilidad de los componentes
	Calidad	- Cantidad de defectos sin corregir - Costo de corrección de defectos - Eficacia en la eliminación de defectos - Cantidad de veces que un módulo fue probado - Tamaño del módulo - Tiempo promedio de corrección de defectos
	Soporte a los clientes	- Tamaño del back log de defectos - Tiempo de respuesta en atender los defectos - Tiempo de resolución de defectos
	Herramientas	- Soporte de herramientas para procesos propuestos

Estas características, muestran que las métricas de software y de sistemas expertos existentes, tampoco pueden considerarse del todo adecuadas, ya que los parámetros que utilizan estos proyectos son de naturaleza diferentes [48] [49] a los de un proyecto de software tradicional, y no se ajustan a sus particularidades.

En la investigación documental realizada, se han identificado dos métodos que permiten estimar el esfuerzo inicial para desarrollar un proyecto de Explotación de Información: uno orientado a proyectos de tamaño mediano o grande y otro orientado a proyectos pequeños (sub-sección 1). Además, se identificó un modelo de procesos para desarrollar

proyectos de Explotación de Información (sub-sección 2), con énfasis en pequeñas y medianas empresas (PyMEs), y un conjunto de procesos de Explotación de información utilizados durante el desarrollo en la fase de modelado (sub-sección 3). Sin embargo, no se han encontrado métricas específicas aplicables al proceso de desarrollo de un proyecto de este tipo. No obstante, a partir de las métricas existentes de la Ingeniería de Software y la Ingeniería del Conocimiento, podrían considerarse aquellas que sean adaptables a los proyectos de Explotación de Información (sub-sección 4).

1. Estimación del Esfuerzo en Proyectos de Explotación de Información

En el trabajo de [48] se propone un método analítico de estimación para proyectos de explotación de información el cual se denomina "Matemático Paramétrico de Estimación para Proyectos de Data Mining" (en inglés Data Mining COst MOdel, o DMCoMo).

Este método es un modelo de estimación de esfuerzo paramétrico de la familia de COCOMO II, que permite estimar los meses/hombre necesarios para desarrollar un proyecto de explotación de información, desde su concepción hasta su puesta en marcha. Para realizar la estimación, se definieron seis categorías con sus factores de costo relacionados [48] [49], que vinculan las características más importantes de los proyectos de Explotación de Información. Estas categorías y sus factores de costos se indican en la Tabla II. No obstante, en validaciones realizadas por [65] [66] del método DMCoMo sobre casos reales, se determinó que el mismo era aplicable a proyectos de tamaño grande y mediano.

Por otra parte, y teniendo en cuenta que los proyectos de Explotación de Información mayormente son requeridos por empresas PyMEs, en el trabajo de [66] se propone un método de estimación de esfuerzo orientado a las características de estas empresas, encuadrándose en el contexto de los proyectos de tamaño pequeño.

Este nuevo método de estimación toma de referencia el método DMCoMo pero con una menor cantidad de factores de costo. Las categorías definidas en el método de estimación para proyectos de Explotación de Información para PyMEs y sus factores de costos se indican en la Tabla III.

Las características consideradas por ambos métodos en la estimación del esfuerzo de un proyecto de Explotación de Información, pueden servir de referencia para proponer una clasificación de métricas significativas aplicables a este tipo de proyectos.

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

Las etapas de desarrollo de los proyectos de Explotación de Información no coinciden naturalmente con las fases mediante las cuales se desarrollan los proyectos de software tradicionales [85], ya que estas etapas están completamente relacionadas con las distintas transformaciones que sufren los datos a lo largo del desarrollo del proyecto. En este sentido, el Modelo de Procesos para Proyectos de Explotación de Información propuesto por [85] plantea dos procesos principales: uno vinculado a la administración de proyectos de explotación de información y otro relacionado con el desarrollo del mismo.

Para el interés de este trabajo, nos centraremos en el Modelo de Proceso de Desarrollo, cuyos subprocesos y tareas fueron definidas a partir de las fases de desarrollo planteadas por la metodología CRISP-DM [9].

TABLA II. CATEGORÍAS Y FACTORES DE COSTO DEL MÉTODO DE ESTIMACIÓN DE ESFUERZO DMCOMO

Categorías	Factores de Costo
Datos	Cantidad de Tablas Cantidad de Tuplas de las Tablas Cantidad de Atributos de las Tablas Grado de Dispersión de los Datos Porcentaje de valores Nulos Grado de Documentación de las Fuentes de Información Grado de Integración de Datos Externos
Modelos	Cantidad de Modelos a ser Creados Tipo de Modelos a ser Creados Cantidad de Tuplas de los Modelos Cantidad y Tipo de Atributos por cada Modelo Cantidad de Técnicas Disponibles para cada Modelo
Plataforma	Cantidad y Tipo de Fuentes de Información Disponibles Distancia y Medio de Comunicación entre Servidores de Datos
Técnicas y Herramientas	Herramientas disponibles para ser usadas Grado de Compatibilidad de las Herramientas con Otros Software Nivel de Formación de los Usuarios en las Herramientas
Proyecto	Cantidad de Departamentos Involucrados en el Proyecto Grado de Documentación que es necesario generar Cantidad de Sitios donde se realizará el Desarrollo y su Grado de Comunicación
Equipo de Trabajo	Grado de Familiaridad con el Tipo de Problema Grado de Conocimiento de los Datos Actitud de los Directivos

TABLA III. CATEGORÍAS Y FACTORES DE COSTO DEL MÉTODO DE ESTIMACIÓN DE ESFUERZO PARA PYMES

Categorías	Factores de Costo
Datos	Cantidad y Tipo de los Repositorios de Datos Disponibles Cantidad de Tuplas Disponibles en la Tabla Principal Cantidad de Tuplas Disponibles en Tablas Auxiliares Nivel de Conocimiento sobre los Datos
Proyecto	Tipo de Objetivo de Explotación de Información Grado de Apoyo de los Miembros de la Organización
Recursos	Nivel de Conocimiento y Experiencia del Equipo de Trabajo Funcionalidad de las Herramientas Disponibles

En la Fig. 1 se observan los subprocesos que componen el Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información definido por [85], en el orden secuencial natural de los mismos. Mientras que en la Tabla IV se muestran las tareas y salidas asociadas a cada uno de estos subprocesos.

Fig. 1. Modelo Generado para el Desarrollo del Proyecto

A continuación se describen cada uno de los subprocesos definidos por [85]:

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

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

El subproceso de *Preparación de los Datos* cubre todas las actividades para construir el conjunto de datos final desde los datos iniciales. En este caso, se toma la información disponible para su manipulación (selección de tablas, atributos y

registros), transformación (limpieza de datos, cambios de formato, construcción de atributos adicionales), y presentación (integración de los datos necesarios en una única tabla), con el objetivo de efectuar su procesamiento a través de técnicas de minería de datos. Las tareas de este subproceso pueden ser realizadas muchas veces y sin un orden preestablecido.

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

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

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

Claramente se observa que los subprocesos y tareas indicados en la Tabla IV difieren de las etapas definidas para un proyecto de desarrollo de software tradicional (inicio, requerimientos, análisis y diseño, construcción, integración y pruebas y cierre). No obstante, si bien este proceso define la manera de desarrollar en forma exitosa un proyecto de Explotación de Información, no especifica qué métricas utilizar para evaluar la calidad del proceso, del producto y de los resultados obtenidos.

Por consiguiente, estos subprocesos y tareas podrían ser una guía para proponer un conjunto de métricas significativas aplicables al desarrollo de un proyecto de Explotación de Información, con énfasis en las características de las empresas PyMEs.

3. Procesos de Explotación de Información

En el trabajo realizado por [5] se definen cinco procesos de Explotación de Información que pueden ser considerados dentro de la etapa de Modelado del desarrollo de un proyecto.

Los procesos de explotación de información definidos son los siguientes:

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

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

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

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

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

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

Para cada uno de los procesos mencionados anteriormente, se propone la utilización de distintas tecnologías, en su mayoría provenientes del campo del aprendizaje automático [23]. No obstante, los procesos son independientes de la tecnología que se utilice para resolverlos.

4. Métricas Existentes aplicables a Proyectos de Explotación de Información

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

La Ingeniería del Software utiliza a los Puntos de Función [3] como técnica para medir el tamaño del proyecto. Esta técnica considera la estimación empírica [14] dada por la relación entre el esfuerzo requerido para construir el software y las características identificadas del mismo, tales como entradas externas (atributos/campos), archivos de interface, salidas, consultas y archivos lógicos internos (tablas).

Las cantidades para cada una de estas características son ajustadas a través de la ponderación y de factores de complejidad para obtener un tamaño expresado en puntos de función.

TABLA IV. TAREAS VINCULADAS AL MODELO DE PROCESO DE DESARROLLO PARA PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

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

En un proyecto de Explotación de Información, el tamaño no se mide a través de puntos de función sino que se establecen tres rangos de tamaño de proyectos (grandes, medianos y pequeños) [65], los cuales se determinan a partir de las características y los valores de los factores de costo considerados por el método DMCoMo [48] [49]. No obstante, el número de tablas y número de atributos podrían ser métricas a tener en cuenta ya que representan las fuentes de datos necesarias para desarrollar el proyecto y la disponibilidad de una cantidad suficiente de datos para aplicar explotación de información.

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

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

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

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

- Número de Atributos: esta métrica permitiría conocer si se dispone de una cantidad suficiente de datos para

aplicar explotación de información, tal como se mencionó anteriormente.

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

III. Descripción del Problema

En este apartado se desarrolla la problemática que intenta solucionar esta investigación (sección *A*) junto con la justificación de las decisiones tomadas para llevar a cabo su resolución (sección *B*) finalizando con las preguntas de investigación que se intenta responder mediante este trabajo de investigación (sección *C*).

A. Identificación del Problema de Investigación

La gestión de un proyecto de software es el primer nivel del proceso de Ingeniería de Software, porque cubre todo el proceso de desarrollo. Para alcanzar el éxito del proyecto se debe comprender el ámbito del trabajo a realizar, los riesgos en los que se puede incurrir, los recursos requeridos, las tareas a llevar a cabo y el plan a seguir. Por consiguiente, la planificación es una de las actividades más importantes del proceso de gestión de un proyecto de software, ya que contempla la correcta estimación del esfuerzo requerido para realizar el proyecto, la duración cronológica del mismo y su costo.

El uso de métricas es una característica importante de todas las disciplinas de ingeniería, entre las que se incluye la Ingeniería de Proyectos de Explotación de Información. Dentro de un marco de trabajo ingenieril, las métricas permiten cuantificar aspectos específicos de un proceso, de un producto o de un proyecto [63]. En este sentido, recolectar métricas constituye el primer paso para saber cómo controlar y mejorar el proceso de desarrollo de software [54].

En el ámbito de la Ingeniería de Software y la Ingeniería del Conocimiento, los proyectos de construcción de software aplican procesos de desarrollo siguiendo un conjunto de tareas específicas. Estos procesos están condicionados al tipo de producto que se desarrolla, lo cual implica que no existe un proceso único que sea aplicable al desarrollo de cualquier proyecto de software. A su vez, para estos procesos se han definido métricas que cubren diferentes enfoques del software y que han sido validadas y utilizadas ampliamente en proyectos de desarrollo de software tradicional y de sistemas expertos.

En el campo de la Ingeniería de Proyectos de Explotación de Información, los proyectos presentan características diferentes respecto a los proyectos de desarrollo de software tradicional y esa diferencia está en la naturaleza del producto resultante [48] [49]. Esto implica que las etapas habituales y métricas definidas para un proyecto de desarrollo clásico no sean apropiadas para un proyecto de Explotación Información [85].

Por otro lado, existen metodologías que acompañan el desarrollo de proyectos de Explotación de Información, y si bien fueron probadas y tienen un buen nivel de madurez en cuanto al desarrollo del proyecto dejan de lado aspectos a nivel gestión de los proyectos [85]. En ese contexto, [85] desarrolló un Modelo de Procesos para Proyectos de Explotación de Información definiendo dos procesos bien diferenciados para desarrollo y gestión. Sin embargo, este modelo carece de métricas para evaluar, controlar y asegurar la calidad del proceso aplicable al desarrollo de un proyecto de Explotación de Información.

B. Descripción del Problema

Las etapas de desarrollo de los Proyectos de Explotación de Información no coinciden naturalmente con las fases mediante las cuales se desarrollan los proyectos de software tradicionales, ya que estas etapas están completamente relacionadas con las distintas transformaciones que sufren los datos a lo largo del desarrollo del proyecto. En consecuencia, [84] propone un Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información.

Por otra parte, [48] define el método DMCoMo como técnica para estimar el esfuerzo al inicio de un proyecto de explotación de información definiendo los factores de costo considerados por el método. No obstante, las conclusiones del trabajo de [49] señalan que este método de estimación es confiable, cuando se lo utiliza para estimar el esfuerzo en proyectos de explotación de información que se encuentran en el rango de esfuerzo de 90 a 185 meses/hombre (es decir, 7,5 a 15,42 años/hombre). Estos proyectos se encuadran dentro de la clasificación de proyectos de tamaño grande [64]. Sin embargo en el trabajo de [65] [66], se señala que también puede aplicarse para proyectos de tamaño mediano. Sin embargo, no se recomienda para proyectos pequeños dado que el método sobreestima el esfuerzo necesario. Debido a esta falencia, en [66] se propone un método específico de estimación de esfuerzo para proyectos pequeños de explotación de información, considerando características particulares de las empresas PyMEs y definiendo los factores de costo adecuados para este otro método.

De esta manera, se plantean dos métodos de estimación inicial de esfuerzo, con características y factores de costo específicos de acuerdo al tamaño del proyecto, y que permiten comparar sus valores con el esfuerzo real aplicado al finalizar cada fase del proceso de desarrollo.

En la Ingeniería de Software y la Ingeniería del Conocimiento, todo proyecto tradicional de construcción de software sigue un proceso de desarrollo que incluye a las métricas como mecanismo de aseguramiento de la calidad (SQA).

Como se mencionó anteriormente, los proyectos de Explotación de Información, también deben aplicar un proceso de desarrollo y utilizar métricas para evaluar distintos aspectos del desarrollo del proyecto y el cumplimiento de los criterios de éxito del problema de negocio al finalizar el mismo. En este sentido, el Modelo de Proceso de Desarrollo definido por [85] constituyó una de las herramientas propuestas y desarrolladas para el campo de la Ingeniería de Proyectos de Explotación de Información. Sin embargo, la necesidad de disponer de métricas significativas para este proceso de desarrollo continúa siendo un problema abierto dentro del cuerpo de conocimiento de esta disciplina.

Considerando el problema planteado, esta investigación se orienta a proponer un conjunto de métricas específicas aplicables al desarrollo de proyectos de Explotación de Información. En particular, existe un interés en los pequeños y

medianos emprendimientos, ya que es en este ámbito donde mayormente se desarrolla este tipo de proyectos.

C. Preguntas de Investigación

Teniendo en cuenta la problemática planteada surgen los siguientes interrogantes que se intentan responder a lo largo de este trabajo de investigación:

- ¿Es posible categorizar e identificar métricas específicas que puedan aplicarse al proceso de desarrollo de un proyecto de Explotación de Información, con énfasis en proyectos de empresas PyMEs?
- ¿Pueden las métricas propuestas para el subproceso de Modelado del proyecto ser aplicables si se consideran los procesos de explotación de información definidos por [5], y que utilizan tecnologías de sistemas inteligentes?
- ¿Es posible medir el éxito del proceso de desarrollo de un proyecto de Explotación de Información, a partir de los objetivos de éxito definidos al inicio del mismo?

Se propone una solución a los interrogantes planteados y el estudio de las métricas por medio de un método empírico, en los próximos apartados.

IV. Solución Propuesta

En este apartado se desarrolla la solución propuesta para la problemática mencionada en el apartado anterior, realizando una descripción general de la propuesta (sección *A*) y la solución a la misma a través de tablas que contienen las métricas definidas (sección *B*).

A. Descripción General de la Propuesta

En base a las características y factores de costo que [48] define para el modelo de estimación DMCoMo, las consideraciones mencionadas en el trabajo de [49] para este método y los factores de costo definidos para el Método de Estimación de Esfuerzo propuesto para PyMEs en [66], se plantea la utilización de aquellos factores que sean aplicables al Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información definido por [85], como forma de clasificación de las métricas propuestas. Este modelo de proceso de desarrollo, con sus subprocesos y tareas, se focaliza en las características particulares de las empresas PyMEs, donde mayormente ocurren los proyectos de Explotación de Información, cuyos parámetros fueron establecidos por [65] en la clasificación de proyectos de tamaño pequeño.

Por otra parte, se plantea como solución que las métricas propuestas para un proyecto de Explotación de Información, consideren la aplicación de los procesos de explotación de información definidos por [5] en la fase de modelado del proyecto.

B. Solución basada en la Propuesta

Se propone como clasificación a utilizar para las métricas propuestas las categorías de Datos (sub-sección *1*), Modelos (sub-sección *2*) y Proyectos (sub-sección *3*), las cuales se consideran que abarcarán todo el modelo de proceso de desarrollo indicado anteriormente.

En la categoría de Datos se incluyen las métricas para los subprocesos de Entendimiento de Datos y Preparación de Datos. En la categoría de Modelos se incluyen las métricas del subproceso de Modelado, y de acuerdo a su tarea de descubrimiento, se proponen los siguientes tipos de modelado para clasificar las métricas: Modelo de Descubrimiento de Grupos, Modelo de Descubrimiento de Reglas y Modelo de Descubrimiento de Dependencias Significativas. Por último, la categoría de Proyectos incluye las métricas para los subprocesos de Evaluación del proceso de desarrollo y la Entrega del proyecto. En la Fig. 2 se muestra la clasificación para las métricas, de acuerdo a las categorías propuestas para un proyecto de Explotación de Información.

Las categorías propuestas en los párrafos precedentes constituyen una aportación de esta investigación, ya que no se encuentran establecidas en el Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información definido por [85].

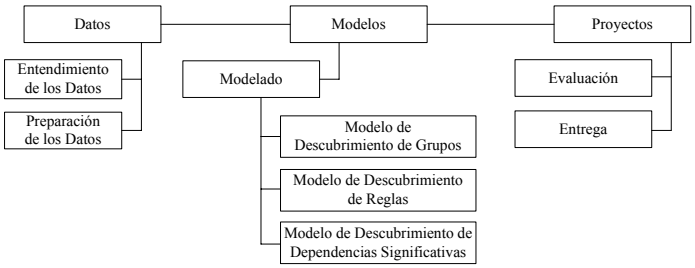

Fig. 2. Clasificación de las Métricas Propuestas

Por otra parte, al tratarse de un modelo de proceso de desarrollo relativamente nuevo, no existen métricas significativas para medir y controlar el progreso del proyecto, evaluar la calidad del proceso, los productos obtenidos y los resultados alcanzados, siguiendo este proceso de desarrollo. En consecuencia, las métricas propuestas constituyen el mayor aporte de esta investigación, por tratarse de métricas específicas que no se encuentran definidas para ningún proyecto de Explotación de Información que aplique este proceso de desarrollo. A su vez, porque las métricas definidas para el subproceso de Modelado, contemplan los procesos de explotación de información definidos por [5] basados en la utilización de sistemas inteligentes [23].

1. Propuesta de Métricas de Datos

Cuando se va a desarrollar un proyecto de Explotación de Información, lo primero que se debe considerar es el volumen de datos. Por ello, resulta fundamental analizar la cantidad y tipo de bases de datos que se necesitan utilizar. Las métricas de datos propuestas en esta sección, hacen referencia al volumen y calidad de los datos a utilizar en el proyecto de Explotación de Información.

A partir de la división de subprocesos definidos en el modelo de proceso de desarrollo para proyectos de Explotación de Información [85], se han propuesto un conjunto de métricas clasificadas en dos categorías: Métricas para el Entendimiento de los Datos (sub-sección *a*) y Métricas para la Preparación de los Datos (sub-sección *b*).

a) Métricas para Entendimiento de los Datos

Una vez identificado los objetivos del proyecto, los criterios de éxito del problema de negocio y se recolectaron las fuentes de datos iniciales, se procede a analizar y conocer cómo están constituidos los datos y a evaluar algunos aspectos de la calidad de los mismos. Por ello, resulta importante medir estos aspectos al inicio del proyecto y comprender el significado de los problemas de calidad que pueden presentar los datos. A partir de modelo de proceso de desarrollo establecido, se han considerado métricas de Entendimiento de los Datos para las tareas que se indican en las Tablas V, VI y VII. Las métricas están ordenadas alfabéticamente.

b) Métricas para Preparación de los Datos

Una vez analizada la calidad de los datos iniciales de las diferentes fuentes de datos, es necesario seleccionar qué atributos y registros son de utilidad para el proyecto de Explotación de Información. Esto involucra analizar y evaluar qué transformaciones, limpieza y construcción de datos se deben efectuar sobre los mismos, a fin de integrarlos en una única tabla, y que sirvan de base para aplicar los modelos de descubrimiento necesarios para el problema de Explotación de Información. A partir del modelo de proceso de desarrollo establecido, se han considerado métricas de Preparación de los Datos para las tareas que se indican en las Tablas VIII, IX, X y XI. Las métricas están ordenadas alfabéticamente.

2. Propuesta de Métricas de Modelos

En los proyectos de Explotación de Información es necesario evaluar la calidad de los modelos obtenidos de la manera más precisa 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 la división de subprocesos definidos en el modelo de proceso de desarrollo para proyectos de Explotación de Información [84], se han propuesto un conjunto de métricas de Modelado (sub-sección *a*), las cuales se complementan con otras métricas, según la tarea de descubrimiento que se aplique al modelo (sub-secciones *b, c y d*).

a) Métricas para Modelado

Una vez definidos los datos con los que se desarrollará el proyecto de Explotación de Información, se deben construir los modelos correspondientes, seleccionando los tipos de modelado de explotación de información, según su tarea de descubrimiento, y calibrando los parámetros en forma adecuada. Dado que pueden aplicarse varios tipos de modelado para generar los modelos del proyecto, y cada uno de ellos tener requisitos específicos sobre los datos, se debe medir y evaluar cada modelo construido, de manera de asegurar que la solución obtenida resuelva eficientemente los objetivos de explotación de información. A partir del modelo de proceso de desarrollo establecido, se han considerado métricas generales de Modelado para las tareas que se indican en las Tablas XII y XIII. Estas métricas están ordenadas alfabéticamente y son aplicables a cualquier tipo de modelado que se necesite utilizar para el proyecto.

En [5] se han definido cinco procesos de Explotación de Información que pueden aplicarse al subproceso de Modelado [85] dentro del proceso de desarrollo. Según su tarea de descubrimiento y aplicación en un proyecto de Explotación de Información, se proponen las siguientes tipos de modelado para clasificar las métricas: Modelos de Descubrimiento de Grupos (sub-sección *b*), Modelos de Descubrimiento de Reglas (sub-sección *c*) y Modelos de Descubrimiento de Dependencias Significativas (sub-sección *c*). A continuación se realiza una breve descripción de cada una de estos tipos de modelos y se presentan las métricas propuestas para su evaluación.

b) Modelo de Descubrimiento de Grupos

El modelo de Descubrimiento de Grupos tiene por objetivo la separación representativa de los datos en grupos o clases, sin ningún criterio de agrupamiento a priori, basándose en la similitud de los valores de sus atributos. Todos los datos de un mismo grupo deben tener características comunes pero a su vez entre los grupos los objetos deben ser diferentes [5].

Dentro de las tecnologías inteligentes [23] apropiadas para realizar agrupamiento están los Mapas Auto-Organizados de Kohonen (SOM – Self-Organized Map, por sus siglas en inglés) [41], algoritmo K-Means [87], entre otros. Al construir un modelo de descubrimiento de grupos, se define un número de grupos evaluando diferentes topologías para escoger de entre todas la más óptima para la solución del problema. El factor de calidad del modelo generado está basado en el número de grupos que se definen, 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 datosPropuesta de Métricas de Datos. A partir del modelo de proceso de desarrollo establecido, se han considerado métricas para la tarea de evaluación del modelo de Descubrimiento de Grupos a las indicadas en la Tabla XIV. Las métricas están ordenadas alfabéticamente.

c) Modelo de Descubrimiento de Reglas

El modelo de Descubrimiento de Reglas es uno de los modelos más importantes de la explotación de información. Se utiliza para encontrar las reglas de asociación y clasificación de un conjunto de casos con base en los valores de sus atributos y su atributo clase [10]. El objetivo es obtener un conjunto potencialmente útil de reglas del tipo *Si <antecendente> Entonces <consecuente>*, que determinen correctamente la clase ante casos no previstos anteriormente [5]. El antecedente de la regla representa las condiciones que deben observarse para que la regla sea aplicable, y el consecuente, la clase que se identifica con la aplicación de la regla. En proyectos de explotación de información, el consecuente de una regla puede referirse también a un grupo de datos, obtenido por medio de un modelo de descubrimiento de grupos [5].

Para encontrar este conjunto de reglas útiles, se necesitan métricas que permitan evaluar las reglas descubiertas sobre un conjunto de casos, pasando de un número muy alto de reglas a un número reducido y realmente útiles. Para ello, se pueden aplicar diversas métricas que se han utilizado ampliamente en otras áreas de investigación como aprendizaje de máquinas (Machine Learning, ML) y minería de datos [73]. En [18] [19], se menciona que un aspecto importante de la minería de datos, es que el conocimiento descubierto debe satisfacer los criterios de precisión (o confianza) y cobertura (o soporte) [2] [44] [45] de la regla. Además de los criterios indicados, este conocimiento debe ser de interés [38] para el proyecto, en el sentido de ser útil y novedoso. El interés se puede medir a través de la cuantificación de los criterios mencionados anteriormente [82], lo que permite evaluar la calidad de las reglas descubiertas y los patrones de conocimiento encontrados [44] [45]. Los algoritmos de inducción TDIDT (Top Down Induction Decisión Trees) [68], son las herramientas de sistemas inteligentes que se emplean para tareas de clasificación en proyectos de Explotación de Información [23]. A esta familia pertenecen los algoritmos: ID3 [68], C4.5 [69] y C5 [70]. Al momento de aplicar los algoritmos de inducción TDIDT se debe tener en cuenta cómo están distribuidos los casos respecto a la clase o grupo [22]. Puede ocurrir que al no estar balanceadas las clases, los algoritmos estén sesgados a predecir un porcentaje más elevado de la clase más favorecida [22]. Las métricas propuestas para evaluar un modelo de descubrimiento de reglas, se basan en analizar la exactitud, tasa de aciertos/errores y precisión del modelo construido y las características de las reglas que describen la pertenencia a una clase o grupo, a partir de las relaciones existentes entre los valores de sus atributos y el atributo clase.

TABLA V. MÉTRICAS PARA ENTENDIMIENTO DE DATOS – DATOS INICIALES

Tarea: Reunir los datos iniciales		
Métrica	**Significado**	**Cálculo de la Métrica**
NA (T)[1]	Número de atributos [3] [49] de la tabla T. Indica el número inicial de atributos que contienen las tablas iniciales a ser consideradas para el proyecto. Mide que se disponga de una cantidad suficiente de datos para aplicar explotación de información. La unidad de medida de la métrica es *atributos* y toma valores mayores a 0.	
NR (T)[1]	Número de registros [49] de la tabla T. Indica el número de tuplas que contienen las tablas iniciales a ser consideradas para el proyecto. Mide que se disponga de una cantidad suficiente de datos para aplicar explotación de información. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	
NT	Número total de tablas [3] [49]. Indica el número inicial de tablas o fuentes de datos (ficheros, archivos, bases de datos, etc.) necesarias al comienzo del proyecto. Se incluyen tablas internas y externas. La unidad de medida de la métrica es *archivos/tablas* y toma valores mayores a 0.	

Nota [1] – Métrica utilizada en las tareas: Explorar los datos, Verificar la calidad de los datos y Seleccionar los datos

TABLA VI. MÉTRICAS PARA ENTENDIMIENTO DE DATOS – EXPLORACIÓN DE DATOS

Tarea: Explorar los datos		
Métrica	**Significado**	**Cálculo de la Métrica**
DVN (T)	Densidad de valores nulos o faltantes en la tabla T. Identifica la proporción total de valores nulos o faltantes en las tablas iniciales para el proyecto. La unidad de medida de la métrica es % (porcentaje) y toma valores entre 0 y 1.	$DVN(T) = \frac{NVN(T)}{NR(T) * NA(T)}$
NCT (T)	Nivel de compleción de la tabla T. Expresa la relación entre la cantidad de valores nulos o faltantes respecto al total de datos de la tabla. Permite determinar cuán completa está la tabla con datos. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	$NCT(T) = 1 - DVN(T)$
NVN (T)[2]	Número de valores nulos o faltantes en la tabla T. Indica la cantidad de valores nulos o faltantes que tiene la tabla a utilizarse inicialmente para el proyecto. La unidad de medida de la métrica es *datos/valores* y toma valores mayores o iguales a 0.	

Nota [2] – Métrica utilizada en la tarea: Verificar la calidad de los datos

TABLA VII. MÉTRICAS PARA ENTENDIMIENTO DE DATOS – CALIDAD DE DATOS

Tarea: Verificar la calidad de los datos		
Métrica	**Significado**	**Cálculo de la Métrica**
GCD (T)	Grado de corrección de los datos de la tabla T. Expresa la relación entre la cantidad de datos con valores erróneos, fuera de rangos (outliers) o nulos/faltantes y la cantidad de datos totales de la tabla. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1. ▪ Si GCD (T) es ALTO implica que los datos de la tabla son útiles para el proyecto de explotación de información. ▪ Si GCD (T) es BAJO implica que deben analizarse los datos y evaluar la relevancia de la tabla.	$GCD(T) = 1 - \frac{NVE(T) + NVN(T)}{NR(T) * NA(T)}$
NVE (T)	Número de valores erróneos en la tabla T. Indica la cantidad de valores mal ingresados o fuera del rango normal que tiene la tabla, ya sea por atributos o registros. Mide la confiabilidad que los datos son verdaderos dependiendo de su fuente de datos y naturaleza. No se consideran erróneos los valores que deban normalizarse. La unidad de medida de la métrica es *datos/valores* y toma valores mayores o iguales a 0.	

Algunas de estas métricas se definen en base a una matriz, llamada Matriz de Confusión [40], las cuales se obtienen cuando se prueba el modelo sobre un conjunto de datos que no intervinieron en la construcción del modelo.

A partir de los valores obtenidos de esta matriz, se cuantifican las métricas para evaluar la calidad del modelo construido y las reglas descubiertas con los algoritmos de inducción TDIDT. Otras métricas, se definen tomando las consideraciones y métricas propuestas en [27] [44] [45], adecuándolas para un modelo de descubrimiento de reglas y su interpretación para proyectos de explotación de información.

Una matriz de confusión permite conocer la distribución del error cometido por un clasificador a lo largo de las clases del problema. Por simplicidad se definirán las métricas para un problema de dos clases, pero las mismas se pueden extender para N clases [86]. Una matriz de confusión general tiene la estructura que se indica en la Fig 3.

En este caso, se presenta una fila y una columna para cada clase, la fila indica el valor real de la clase asociada al caso evaluado y cada columna el valor clasificado por el modelo para ese mismo caso. Los valores que se encuentran a lo largo de la diagonal principal de la matriz, indicados como NCVA y NCVB, son las clasificaciones correctas del modelo. Los valores de la diagonal secundaria, indicados como NCFA y NCFB, representan los errores (la confusión) entre las clases.

		Clase Clasificada		
		Clase A	***Clase B***	***Total***
Clase Real	***Clase A***	Número de casos clasificados como A y son de clase A (NCVA)	Número de casos clasificados como B pero son de clase A (NCFB)	Total de casos de la clase A
	Clase B	Número de casos clasificados como A pero son de clase B (NCFA)	Número de casos clasificados como B y son de clase B (NCVB)	Total de casos de la clase B
	Total	Total de casos clasificados como clase A	Total de casos clasificados como clase B	Número total de casos (NTC)

Fig. 3. Estructura general de una Matriz de Confusión de dos clases

Se considera que ocurre un error de clasificación en el modelo, cuando un caso es clasificado como clase A cuando es realidad es de clase B, o viceversa [86]. Un buen resultado del modelo se corresponde con valores altos en la diagonal principal y con valores bajos, idealmente cero, en la diagonal secundaria [86].

Cuando el problema presenta N clases de clasificación o predicción, los resultados de la clasificación se expresan en una matriz de confusión de dimensión N x N, como se observa en la Fig. 4.

TABLA VIII. MÉTRICAS PARA PREPARACIÓN DE DATOS – SELECCIÓN DE DATOS

Tarea: Seleccionar los datos		
Métrica	**Significado**	**Cálculo de la Métrica**
NANC (T)[3]	Número de atributos no correctos en la tabla T. Indica la cantidad de atributos de la tabla que no son útiles para el proyecto, por tener muchos errores o valores nulos. La unidad de medida de la métrica es *atributos* y toma valores mayores o iguales a 0.	NA(T) = UTILES(T) + NO _ UTILES(T) donde: UTILES(T) = NASE(T) + NAUD(T) NO _ UTILES(T) = NANC(T) + NANS(T)
NANS (T)[3]	Número de atributos no significativos de la tabla T. Indica la cantidad de atributos de la tabla que no son relevantes para el proyecto. No necesitan ningún tipo de análisis del contenido. La unidad de medida de la métrica es *atributos* y toma valores mayores o iguales a 0.	
NASE (T)[3]	Número de atributos útiles sin errores en la tabla T. Indica la cantidad de atributos cuyos datos tienen ausencia de valores nulos o faltantes, no contienen errores ni requieren normalización, es decir, son datos que no necesitan ningún tipo de corrección. La unidad de medida de la métrica es *atributos* y toma valores mayores o iguales a 0.	
NAUD (T)[3]	Número de atributos útiles pero con defectos en la tabla T. Indica la cantidad de atributos útiles cuyos datos que deben corregirse, completarse o normalizarse con algún formato específico, para ser utilizados en el proyecto. La unidad de medida de la métrica es *atributos* y toma valores mayores o iguales a 0.	
NRDP (T)[3]	Número de registros duplicados en la tabla T. Indica la cantidad total de registros que se identificaron como duplicados en la tabla. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores o iguales a 0.	
NRNC (T)[3]	Número de registros no correctos en la tabla T. Indica la cantidad de registros de la tabla que no son útiles para el proyecto, por tener muchos errores o valores nulos. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores o iguales a 0.	
GUA (T)	Grado de utilidad de los atributos de la tabla T. Expresa la proporción de atributos útiles de la tabla para el proyecto de explotación de información. A cada atributo se le asigna una ponderación, según su nivel de utilidad. ▪ Si GUA (T) es BAJO implica que la mayoría de los atributos de la tabla no son de útiles para el proyecto, pudiéndose descartar la tabla. ▪ Si GUA (T) es MEDIO implica que los atributos de la tabla son aceptablemente usables para el proyecto pero requiere correcciones. ▪ Si GUA (T) es ALTO implica que la mayoría de los atributos de la tabla son de utilidad para el proyecto y no requiere demasiadas correcciones. Cuando GUA (T) es MEDIO o ALTO se deben seleccionar los atributos útiles y efectuar correcciones sobre los atributos útiles con defectos para aumentar la utilidad de los mismos. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	$GUA(T) = \frac{NA(T) - (NO\_UTILES(T) + 0{,}5 * NAUD(T))}{NA(T)}$ Ponderación definida ▪ Atributos útiles con defectos = 0.5 ▪ Atributos No Útiles = 1

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

TABLA IX. MÉTRICAS PARA PREPARACIÓN DE DATOS – LIMPIEZA DE DATOS

Tarea: Limpiar los datos		
Métrica	**Significado**	**Cálculo de la Métrica**
NAU (T)[4]	Número total de atributos que son de utilidad en la tabla T. La unidad de medida de la métrica es *atributos* y toma valores mayores o iguales a 0.	NAU(T) = NASE(T) + NAUD(T)
NANU (T)	Número total de atributos que no son de utilidad en la tabla T. Se incluyen los atributos no significativos, los que no son correctos y los que tienen una importante cantidad de valores nulos o faltantes. La unidad de medida de la métrica es *atributos* y toma valores mayores o iguales a 0.	NANU(T) = NANS(T) + NANC(T)
NRNU (T)	Número de registros que no son de utilidad en la tabla T. Se incluyen también los registros que están duplicados. Estos registros La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores o iguales a 0.	NRNU(T) = NRNC(T) + NRDP(T)
NRU (T)	Número de registros útiles sin errores de la tabla T. Indica la cantidad de registros cuyos datos tienen ausencia de valores nulos o faltantes y no contienen errores, es decir, son datos del registro que no necesitan ningún tipo de corrección. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores o iguales a 0.	NRU(T) = NR(T) – NRNU(T)

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

TABLA X. MÉTRICAS PARA PREPARACIÓN DE DATOS – CONSTRUCCIÓN DE DATOS

Tarea: Construir los datos		
Métrica	**Significado**	**Cálculo de la Métrica**
NANI (TI)[5]	Número de atributos nuevos para integrar a la tabla TI. Indica la cantidad de atributos nuevos que necesitan agregarse o construirse en la tabla única para el proyecto de explotación de información. La unidad de medida de la métrica es *atributos* y toma valores mayores o iguales a 0.	

Nota [5] – Métrica utilizada en la tarea: Integrar los datos

TABLA XI. MÉTRICAS PARA PREPARACIÓN DE DATOS – INTEGRACIÓN DE DATOS

Tarea: Integrar los datos		
Métrica	**Significado**	**Cálculo de la Métrica**
NA (TI)	Número de atributos de la tabla integrada TI. Indica la cantidad total de atributos que se utilizarán para el proyecto y mide que se disponga de una cantidad suficiente de datos para aplicar explotación de información. Representa la suma de todos los atributos útiles por cada tabla analizada, agregándole además los atributos nuevos que se creen para el proyecto. La unidad de medida de la métrica es *atributos* y toma valores mayores o iguales a 0.	$NA(TI) = NANI(TI) + \sum_{i} NAU(T_i)$ donde *i* representa aquellas tablas cuyo grado de utilidad de la métrica GUA (T) es MEDIO O ALTO.
NR (TI)	Número de registros de la tabla integrada TI. Indica la cantidad total de tuplas que se utilizarán para el proyecto y mide que se disponga de una cantidad suficiente de datos para aplicar explotación de información. Representa la suma de todos los registros útiles por cada tabla analizada. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores o iguales a 0.	

TABLA XII. MÉTRICAS PARA MODELOS – GENERACIÓN EL DISEÑO DEL TEST

Tarea: Generar el diseño del test		
Métrica	**Significado**	**Cálculo de la Métrica**
NCE (C)[7]	Número de casos del conjunto de entrenamiento distribuido por cada clase o grupo. Esta métrica permite conocer cómo es la proporción de casos seleccionados por clase o grupo para entrenar el modelo, a fin de evitar que el modelo tienda a clasificar mejor los casos de la clase mayoritaria. La unidad de medida de la métrica es *% (porcentaje)* y toma valores mayores o iguales a 0.	
NCE (M)[6,7]	Número de casos a utilizar para el entrenamiento del modelo M. Indica la cantidad de casos que se utilizarán como conjunto de entrenamiento del modelo. Como heurística, este número suele representar los dos tercios del conjunto total de casos [23]. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	$NCE(M) = \sum_{i} NCE(C_i)$ donde el subíndice *i* representa cada clase o grupo.
NCP (C)[7]	Número de casos del conjunto de prueba distribuido por cada clase o grupo. Esta métrica permite conocer cómo es la proporción de casos seleccionados para probar el modelo, a fin de evitar que el modelo tienda a clasificar mejor los casos de la clase mayoritaria. La unidad de medida de la métrica es *% (porcentaje)* y toma valores mayores o iguales a 0.	
NCP (M)[6,7]	Número de casos a utilizar para las pruebas del modelo M. Indica la cantidad de casos que se utilizarán para validar la calidad del modelo construido y para chequear cada objetivo de explotación de información, en forma separada. Como heurística, el número de casos suele representar un tercio del conjunto total de casos [23]. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	$NCP(M) = \sum_{i} NCP(C_i)$ donde el subíndice *i* representa cada clase o grupo.

Nota [6] – Métrica utilizada en las tareas: Construir el Modelo, Evaluar el modelo de Descubrimiento de Reglas / Dependencias Significativas

Nota [7] – Métrica utilizada en las tareas: Evaluar el modelo de Descubrimiento de Reglas / Dependencias Significativas

TABLA XIII. MÉTRICAS PARA MODELOS – CONSTRUCCIÓN DEL MODELO

Tarea: Construir el modelo		
Métrica	**Significado**	**Cálculo de la Métrica**
NA (M)[8]	Número de atributos del modelo M [49]. Indica la cantidad de atributos a utilizar para construir cada modelo. La unidad de medida de la métrica es *atributos* y toma valores mayores a 0. A mayor número de atributos, mayor es el tiempo requerido para la construcción del modelo [48].	
NCL (M)	Número de clases a utilizar en el modelo M. Indica la cantidad de clases instanciadas de un atributo clase que se utilizan para tareas de clasificación de clases. La unidad de medida de la métrica es *clases/grupos* y toma valores mayores a 0.	
NMOD[9]	Número de modelos [49] a construir para el proyecto de explotación de información. Indica la cantidad de modelos que se deben construir para satisfacer los objetivos de explotación de información planteados para el proyecto. La unidad de medida de la métrica es *modelos* y toma valores mayores a 0.	
NTC (M)	Número total de casos del modelo M [49]. Indica la cantidad total de registros que se utilizan para construir cada modelo. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	$NTC(M) = NCE(M) + NCP(M)$

Nota [8] – Métrica utilizada en la tarea: Evaluar el modelo de Descubrimiento de Reglas

Nota [9] – Métrica utilizada en la tarea: Evaluar los resultados

TABLA XIV. MÉTRICAS PARA MODELOS – EVALUACIÓN DEL MODELO DE DESCUBRIMIENTO DE GRUPOS

Tarea: Evaluar el modelo		
Métrica	**Significado**	**Cálculo de la Métrica**
NCG (G)	Número de casos asignados a cada grupo G. Esta métrica permite conocer la distribución de los casos, agrupados por cada grupo identificado. La suma de los casos asignados en cada grupo debe coincidir con el número total de casos del modelo. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	$NTC(M) = \sum_{i=1}^{NGR(M)} NCG(G_i)$
NGR (M)	Número de grupos generados por el modelo M. La unidad de medida de la métrica es *grupos* y toma valores mayores a 0.	

Sin embargo, y como se muestra en el ejemplo de la Fig. 5, este tipo de problemas se puede reducir a problemas de clasificación de dos clases, si cada una se considera de forma separada frente a la unión del resto de las clases, obteniendo por lo tanto N matrices de confusión [57].

		Clase Clasificada				
		Clase C_1	***Clase C_2***	**…**	***Clase C_N***	***Total***
Clase real	***Clase C_1***	nc_{11}	nc_{12}	…	nc_{1N}	TRC_1
	Clase C_2	nc_{21}	nc_{22}	…	nc_{2N}	TRC_2
	⋮	…	…	…	…	…
	Clase C_N	nc_{N1}	nc_{N2}	…	nc_{NN}	TRC_N
	Total	TCC_1	TCC_2	…	TCC_N	NTC

Fig. 4. Estructura general de una Matriz de Confusión de N clases

donde:

- N es el número de clases
- ncij representa el número de casos de la clase Ci clasificados como Cj
- TRCi es el número total de casos reales de la clase Ci
- TCCi es el número total de casos clasificados como clase Ci
- NTC es el número total de casos

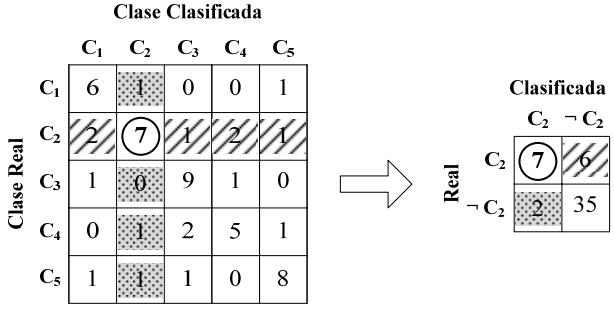

Fig. 5. Reducción de una Matriz de Confusión de 5 clases a una de 2 clases

A partir del modelo de proceso de desarrollo establecido, se han considerado métricas para la tarea de evaluación del modelo de Descubrimiento de Reglas a las indicadas en la Tabla XV. Las métricas están ordenadas alfabéticamente.

Estas métricas se basan en la capacidad que tiene el modelo construido para encontrar las reglas de comportamiento o de pertenencia que determinen correctamente la clase o grupo asociado a un caso de prueba o a nuevos casos. No obstante, la decisión sobre qué modelo de descubrimiento de reglas resulta más adecuada para el proyecto de Explotación de Información, es siempre algo más complicado ya que intervienen otros aspectos, como son los riesgos que suponen los clasificadores empleados y la valoración de las consecuencias que conllevan las clasificaciones incorrectas en el contexto del proyecto.

Además de las métricas propuestas existen otras (curva ROC, curva LIFT, entropía cruzada, coeficiente de Kappa, etc.) que podrían ser utilizadas para evaluar la calidad de un modelo de clasificación [8] [86], pero su complejidad y utilidad van más allá del objetivo de este trabajo de investigación.

d) Modelo de Descubrimiento de Dependencias Significativas

El modelo de Descubrimiento de Dependencias Significativas consiste en encontrar modelos que describan dependencias o asociaciones significativas entre los datos [5]. Esto implica identificar y ponderar las características o factores que tienen mayor incidencia sobre un determinado resultado de un problema [5]. Al igual que con el modelo de descubrimiento de reglas, este modelo también se relaciona con tareas de clasificación y predicción. En este caso, las dependencias pueden ser vistas como los valores que deben tomar determinados atributos del modelo construido, teniendo la información de los otros atributos, para que los mismos resulten significativos en la clasificación o predicción de una clase o grupo [5].

Dentro de las técnicas de sistemas inteligentes apropiadas para realizar descubrimiento de dependencias significativas en un proyecto de Explotación de Información, se encuentran las Redes Bayesianas [23]. Estas redes pueden ser aplicadas para identificar atributos significativos en grandes masas de información [7], ponderar reglas de comportamiento o de pertenencia a grupos [6], y detectar patrones de comportamiento en análisis de series temporales [28].

Un modelo de descubrimiento de dependencias significativas que utiliza Redes Bayesianas puede verse como un modelo de clasificación de clases, en la cual se asume que las relaciones de dependencias entre los atributos del conjunto de datos son condicionalmente independientes entre sí dado un atributo clase [20] [42].

Por tratarse de un modelo de clasificación, pueden aplicarse las métricas de exactitud EXCT (M), tasa de aciertos TAM (C), tasa de errores TEM (C) y precisión del modelo PDM (C), que se definieron en la Tabla XV para el modelo de Descubrimiento de Reglas. Estas métricas surgen de la matriz de confusión del modelo construido para un problema de N clases. En este caso, los valores a predecir o clasificar son los correspondientes al atributo clase considerado en el modelo de dependencias. A su vez, se proponen otras métricas específicas para el modelo de descubrimiento de dependencias significativas.

A partir del modelo de proceso de desarrollo establecido, se han considerado métricas para la tarea de evaluación del modelo de Descubrimiento de Dependencias Significativas a las indicadas en la Tabla XVI. Las métricas están ordenadas alfabéticamente.

Todos los tipos de modelos descriptos para el subproceso Modelado [85] y las métricas propuestas para su evaluación dentro de un proyecto, permiten aplicar los procesos de explotación definidos por [5]. 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.

3. Propuesta de Métricas de Proyectos

Si bien en [48] [49] [65] [66] se plantean los factores de costo que involucran las características más importantes de los proyectos de Explotación de Información, dentro de la cual se encuentra la categoría de Proyecto, no se hace ninguna mención sobre los criterios para evaluar los resultados obtenidos en el proceso de desarrollo y el tiempo insumido por cada tarea vinculada al mismo. Una vez construidos los modelos necesarios para el proyecto de Explotación de Información, evaluada su calidad y los patrones de conocimiento descubiertos, se debe analizar e interpretar si los resultados obtenidos son útiles y si cumplen con los objetivos planteados al inicio del proyecto y los criterios de éxito del negocio.

TABLA XV. MÉTRICAS PARA MODELOS – EVALUACIÓN DEL MODELO DE DESCUBRIMIENTO DE REGLAS

Tarea: Evaluar el modelo		
Métrica	**Significado**	**Cálculo de la Métrica**
COBER (R)	Cobertura o soporte de la regla [2] [44] [45] de pertenencia a una clase o grupo. Mide la proporción de casos de entrenamiento a los que se le puede aplicar cada regla del modelo. Cuanto mayor sea este valor, mayor calidad y utilidad tiene la regla para asociar casos a una clase particular. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	$COBER(R) = \frac{NCCNS(R)}{NCE(C)}$
EXCT (M)[10]	Exactitud [40] del modelo. Mide la proporción de casos clasificados correctamente por el modelo respecto del número total de casos utilizados. Esta métrica se aplica tanto para los casos del conjunto de entrenamiento del modelo como para los casos de prueba y permite evaluar la capacidad del modelo construido para predecir y clasificar nuevos casos dentro del dominio del proyecto. ▪ Si EXCT (M) es ALTO implica que el modelo es capaz de clasificar correctamente los casos que se le presenten del dominio y de encontrar reglas que describen el comportamiento y pertenencia a cada clase o grupo. ▪ Si EXCT (M) es BAJO implica que el modelo no es capaz de clasificar correctamente los casos que se le presenten del dominio, lo cual implica que no es un modelo confiable. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	En fase de entrenamiento: $EXCT(M) = \frac{NCVA + NCVB}{NCE(M)}$ En fase de prueba: $EXCT(M) = \frac{NCVA + NCVB}{NCP(M)}$ En modelos construidos para varias clases, la exactitud se calcula sumando los casos de la diagonal principal de la matriz de confusión respecto al número total de casos del modelo.
NAPR (M)	Número de atributos de la precondición de las reglas [27] de pertenencia a una clase o grupo. Representa la cantidad de atributos diferentes utilizados por el modelo de descubrimiento de reglas para aplicar las condiciones de pertenencia a una clase o grupo sobre los casos de entrenamiento. Si un mismo atributo se utiliza en otra regla, no se lo debe contabilizar nuevamente en esta cantidad. No se incluye el atributo clase dentro de este número. La unidad de medida de la métrica es *atributos* y toma valores mayores a 0.	
NCCNS (R)	Número de casos que satisfacen la aplicación de la regla de comportamiento a una clase o de pertenencia a ésta, es decir, la clase indicada como consecuente de la regla. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	
NCFA	Número de casos que fueron incorrectamente clasificados por el modelo como de clase A, pero que pertenecen a otra clase. Mide la cantidad de casos erróneamente clasificados de una clase. Es deseable que el valor de esta métrica converja a 0 para asegurar una alta tasa de aciertos en el modelo. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	
NCFB	Número de casos que pertenecen a la clase A pero que fueron incorrectamente clasificados por el modelo en otra clase. Mide la cantidad de casos erróneamente clasificados de una clase. Para un problema de N clases, representa todos aquellos casos que fueron clasificados incorrectamente en otras clases diferentes de la clase A. Es deseable que el valor de esta métrica converja a 0 para asegurar una alta tasa de aciertos en el modelo. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	
NCPRC (R)	Número de casos que satisfacen la precondición de la regla, independientemente de la clase o grupo a la que pertenece. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	
NCVA	Número de casos pertenecientes a la clase A que fueron correctamente clasificados por el modelo en esa misma clase. Mide la cantidad de casos correctamente clasificados de una clase. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	
NCVB	Número de casos que pertenecen a la clase B que fueron correctamente clasificados por el modelo en esa misma clase. Mide la cantidad de casos correctamente clasificados de una clase. Para un problema de N clases, la clase B representa todas las clases diferentes a la clase A. La unidad de medida de la métrica es *registros/tuplas* y toma valores mayores a 0.	
NRGL(C)	Número de reglas [27] descubiertas por el modelo por cada clase o grupo. Indica la cantidad de reglas que describen el comportamiento y la pertenencia de un conjunto de casos a cada clase o grupo del modelo. ▪ Si NRGL (C) es BAJO resulta fácilmente identificable el comportamiento de la clase o grupo, lográndose una cobertura de casos más alta por cada regla. ▪ Si NRGL (C) es ALTO se dificulta identificar el comportamiento de la clase o grupo, lográndose una cobertura de casos más baja por cada regla. La unidad de medida de la métrica es *reglas de pertenencia/comportamiento* y toma valores mayores a 0.	
PRCR (R)	Precisión o confianza de la regla [2] [44] [45] de asociación a una clase. Mide la probabilidad condicionada de los casos asociados con una regla particular. Representa la proporción de casos que cumplen con la regla de pertenencia a una clase o grupo, respecto del total de casos considerados en la precondición de la misma. Cuanto mayor sea este valor, mayor calidad y confiabilidad tiene la regla para asociar casos a una clase particular. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	$PRCR(R) = \frac{NCCNS(R)}{NCPRC(R)}$
PDM (C)[10]	Precisión o confianza [40] del modelo para clasificar una clase. Representa la proporción de casos que pertenecen a una clase respecto del total de casos clasificados por el modelo para esa misma clase. Mide la efectividad del modelo para clasificar una clase particular. Es deseable que el valor de esta métrica converja a 1, lo que indica que el modelo es capaz de clasificar correctamente nuevos casos a su clase asociada. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	$PDM(C_i) = \frac{NCVC_i}{NCVC_i + NCFC_i}$ donde C_i es la clase correspondiente a la matriz de confusión que se está evaluando.
TAM (C)[10]	Tasa de aciertos del modelo en la clasificación de casos a una clase. Representa la proporción de casos de prueba de una clase que fueron correctamente clasificados por el modelo respecto del total de casos de la misma clase. Mide la eficacia del modelo para clasificar correctamente un nuevo caso en esa misma clase. ▪ Si TAM (C) es ALTO implica que el modelo es bueno para clasificar la clase, es decir que un nuevo caso tiene una alta probabilidad de ser clasificado correctamente. ▪ Si TAM (C) es BAJO implica que el modelo no es bueno para clasificar la clase, es decir que	$TAM(C_i) = \frac{NCVC_i}{NCP(C_i)}$ donde C_i es la clase correspondiente a la matriz de confusión que se está evaluando. En la matriz de confusión se

Tarea: Evaluar el modelo		
Métrica	**Significado**	**Cálculo de la Métrica**
	un nuevo caso tiene una baja probabilidad de ser clasificado correctamente. En este caso se debe evaluar la elección de los casos de entrenamiento y prueba del modelo y repetir las pruebas. Puede suceder que la elección de los casos de entrenamiento no sean suficientemente representativos de los casos de prueba. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	visualiza como: $TAM(C) = \frac{NCVA}{NCVA + NCFB}$
TEM (C)[10]	Tasa de errores del modelo en la clasificación de casos a una clase. Representa la proporción de casos de prueba de una clase que fueron incorrectamente clasificados respecto del total de casos de la misma clase. Es deseable que el valor de la métrica converja a 0 para garantizar que el modelo es capaz de clasificar correctamente nuevos casos a su clase asociada. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	$TEM(C_i) = 1 - TAM(C_i)$ donde C_i es la clase correspondiente a la matriz de confusión y que se está evaluando.
USAT (M)	Usabilidad de los atributos [27] del modelo. Indica la proporción de atributos utilizados por las reglas generadas sobre el conjunto de casos de entrenamiento, que describen la pertenencia a las clases o grupos del modelo, respecto del número total de atributos usados en el diseño del modelo. Es deseable que el valor de esta métrica converja a 1 para garantizar que el modelo necesita todos los atributos para encontrar las reglas de pertenencia a las clases o grupos. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	$USAT(M) = \frac{NAPR(M)}{NA(M)}$

Nota [10] – Métrica utilizada en la tarea: Evaluar el modelo de Descubrimiento de Dependencias Significativas

TABLA XVI. MÉTRICAS PARA MODELOS – EVALUACIÓN DEL MODELO DE DESCUBRIMIENTO DE DEPENDENCIAS SIGNIFICATIVAS

Tarea: Evaluar el modelo		
Métrica	**Significado**	**Cálculo de la Métrica**
GINC (A_{Vi})	Grado de incidencia que cada valor V_i de un atributo significativo A tiene sobre una clase o grupo del atributo clase. Es deseable que esta métrica converja a 1 para obtener aquellas condiciones o factores que ejercen mayor incidencia sobre un determinado resultado en el dominio del proyecto. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	$GINC(A_{vi}) = \frac{NCAT(A_{vi})}{NCE(C)}$ donde A son los atributos significativos que se obtienen de la métrica NAPR (M).
NCAT (A_{Vi})	Número de casos cubiertos por una clase o grupo del atributo clase, cuando un atributo significativo A toma el valor V_i. Indica la proporción de casos de un atributo significativo que incide sobre el valor del atributo clase. La unidad de medida de la métrica es *registros/tuplas* y toma valores entre mayores a 0.	
NVLC (M)	Número de valores distintos que toma el atributo clase en el modelo M. Para cada valor que toma el atributo clase, se analiza el grado de incidencia que tienen los valores del los atributos significativos. Esta métrica representa la dimensión que toma la matriz de confusión para el modelo. La unidad de medida de la métrica es *valor* y toma valores mayores a 0.	

Este análisis, de resultar satisfactorio contribuye a la aprobación del proyecto por parte del cliente/usuario experto del negocio, su entrega final y cierre. Por otra parte, estos resultados determinan si el proceso de construcción del proyecto fue exitoso y si las tareas se realizaron dentro de los plazos estimados al inicio del mismo. Por consiguiente, las métricas que se proponen para evaluar los resultados del proceso, la duración, el esfuerzo real del proyecto y el desvío respecto del esfuerzo estimado, se incluyen dentro de esta categorización.

A partir de la división de subprocesos definidos en el modelo de proceso de desarrollo para proyectos de Explotación de Información [85], se han propuesto un conjunto de métricas clasificadas en dos categorías: métricas orientadas a la evaluación del proceso de desarrollo (sub-sección *a*) y métricas orientadas a la entrega del proyecto de Explotación de Información (sub-sección *b*).

a) Métricas para Evaluación del Proceso de Desarrollo del Proyecto

En la tarea de evaluación de los modelos descripta en la sección anterior, se proponen las métricas de exactitud EXCT (M) y tasa de aciertos TAM (C), como métricas para evaluar la calidad de los modelos desarrollados y probados para el proyecto de Explotación de Información. Estas métricas, son un indicador de la bondad del modelo construido para proporcionar patrones de comportamiento y conocimiento descubierto, a partir del conjunto de datos utilizado en el desarrollo del proyecto. En el conocimiento descubierto existen patrones que son más interesantes que otros [38], razón por la cual, los usuarios expertos en el negocio deben evaluar cuáles de ellos son los que le aportan conocimiento que le sean de utilidad (que se lo pueda utilizar posteriormente) [58]. A su vez, este conocimiento debe ser sorprendente, inesperado y novedoso (conocimiento nuevo para el usuario) [73].

El interés por los resultados obtenidos, es un factor subjetivo pero necesario para un proyecto de explotación de información, ya que depende de lo que el usuario experto considera como conocimiento útil y novedoso [46]. No obstante, se lo puede relacionar con el grado de cumplimiento de los objetivos de explotación de información, definidos al inicio del proyecto (en el subproceso Entendimiento del Negocio [85]), y en consecuencia, con el éxito del proceso de desarrollo.

Para definir la métrica de éxito de los resultados obtenidos en el proceso de desarrollo, se considera como parámetro el nivel de satisfacción e interés del conocimiento que presenta para el usuario cada modelo desarrollado y probado, tomando un criterio de valoración ALTO, MEDIO, BAJO y MUY BAJO. A su vez, cada modelo de explotación de información se pondera de acuerdo a este interés, asignándole los valores de ponderación correspondientes. Para ello, se propone la utilización del Índice Neto de Satisfacción (del inglés Net Satisfaction Index, NSI) considerado en [37].

El Índice Neto de Satisfacción se considera en un rango entre 0% y 100% (representado con valores entre 0 y 1) y distribuido proporcionalmente por nivel de interés de resultados, ponderando con mayor valor los modelos que presenten un conocimiento interesante, útil y novedoso para el proyecto, como se indica en la Tabla XVII.

TABLA XVII. PONDERACIÓN ASOCIADA AL INTERÉS DEL USUARIO SEGÚN LOS RESULTADOS DE LOS MODELOS

Interés de Resultados	Significado	Ponderación
Muy Bajo	Los resultados obtenidos con los modelos no son de interés para el usuario.	0
Bajo	Los resultados obtenidos aportan algo de conocimiento interesante pero es poco útil para el proyecto.	0.3
Medio	Los resultados obtenidos muestran conocimiento interesante para el usuario, pero se necesitan refinar y clarificar los mismos para que sean útiles para el proyecto.	0.6
Alto	Los resultados obtenidos con los modelos aportan conocimiento interesante para el usuario, siendo además útil y novedoso para el proyecto.	1

Cabe aclarar, que independientemente del interés de los resultados obtenidos con los modelos, respecto a los objetivos definidos al inicio, para que un proyecto de explotación de información sea exitoso, deben cumplirse las siguientes condiciones previas: las fuentes de datos o tablas utilizadas están implementadas con tecnologías que permiten un fácil acceso y manipulación (tareas de limpieza, formateo e integración); se cuenta con el apoyo de los principales interesados en el proyecto (stakeholders) - directivos de la organización, gerentes de nivel medio y/los usuarios finales; es posible realizar una planificación correcta del proyecto considerando la realización de buenas prácticas ingenieriles con el tiempo adecuado; y existen personas en el equipo de trabajo con experiencia en proyectos similares [67]. A partir del modelo de proceso de desarrollo establecido, se han considerado métricas de evaluación del proceso de desarrollo del proyecto para la tarea que se indica en la Tabla XVIII. Las métricas están ordenadas alfabéticamente.

b) Métricas para Entrega del Proyecto

Si los modelos generados son exitosos en función de los criterios de éxito establecidos al inicio del proyecto, se procede a la explotación del modelo construido [71]. Para ello, se elabora luego un informe final conteniendo, por cada grupo de usuarios interesados, una valoración de los resultados alcanzados respecto a los objetivos del problema, los descubrimientos obtenidos del proceso de explotación de información y de ser necesario los pasos para implementar un proceso de explotación de información repetible. Este informe puede incluir también una descripción del proceso seguido, cualquier desviación respecto del plan de proyecto original [9], y el registro de métricas que suministren información relevante a tiempo, que permitan establecer objetivos de mejora para el proceso o en los productos obtenidos. Dentro de las métricas a utilizar se destaca la medición de la duración real del proyecto, el esfuerzo realizado para cumplimentar cada una de las tareas del proceso de desarrollo aplicado y la productividad del proyecto [63]. Esto permite comparar los valores reales obtenidos con las estimaciones realizadas al inicio del proyecto e identificar los desvíos ocurridos en el mismo. Por otra parte, permite conocer, en etapas tempranas, el tiempo que insume cada tarea durante el desarrollo del proyecto de explotación de información, tener una aproximación de los tiempos de los otros subprocesos y del esfuerzo global del proyecto [71]. Estas mediciones facilitan el control del progreso del proyecto, permitiendo realizar los ajustes necesarios en caso de identificar retrasos en alguna de las tareas del proceso de desarrollo.

A partir del modelo de proceso de desarrollo establecido, se han considerado métricas para entrega del proyecto para la tarea que se indica en la Tabla XIX.

V. COMPORTAMIENTO DE LAS MÉTRICAS

En este apartado se analiza y estudia el comportamiento de las métricas mencionadas como solución en el apartado anterior. Se comienza con una descripción de las generalidades (sección *A*) del método de estudio a aplicar, luego se mencionan qué métricas se someterán a estudio (sección *B*) por este método y se finaliza analizando el comportamiento de las métricas de Datos (sección *C*), métricas de Modelos (sección *D*) y métricas de Proyectos (sección *E*) que se proponen en este trabajo.

A. Generalidades

En virtud que resulta difícil obtener datos de proyectos de Explotación de Información reales aplicados a empresas PyMEs, y al no existir hasta el momento métricas que permitan analizar el comportamiento de las mismas para estos proyectos, se decide utilizar una técnica de validación empírica por simulación basada en el método de Monte Carlo. El método de Monte Carlo [35] da solución a una gran variedad de problemas matemáticos haciendo experimentos con muestreos estadísticos en una computadora. El método es aplicable a cualquier tipo de problema, ya sea estocástico o determinístico. La simulación de Monte Carlo es una técnica que combina conceptos estadísticos (muestreo aleatorio) con la capacidad que tienen las computadoras para generar números pseudo-aleatorios y automatizar cálculos. La simulación consiste en crear un modelo matemático del sistema, proceso o actividad que se quiere analizar, identificando aquellas variables independientes cuyo comportamiento aleatorio determina el comportamiento de las variables dependientes. Una vez identificadas las variables independientes, se lleva a cabo un experimento que consiste en generar, con ayuda de una computadora, muestras aleatorias (valores concretos) para dichas variables, y analizar el comportamiento de las variables dependientes ante los valores generados. Tras repetir N veces este experimento, se dispone de N observaciones sobre el comportamiento del sistema, lo cual resulta de utilidad para entender el funcionamiento del mismo. El análisis será más preciso cuanto mayor sea el número de experimentos que se lleven a cabo. De esta manera, se decide estudiar el comportamiento de las métricas de Datos, Modelos y Proyectos, propuestas en este trabajo de Investigación, generando diferentes bancos de pruebas simulados de proyectos de explotación de información. Se consideran además, las restricciones establecidas por [65] para un proyecto de tamaño pequeño y de aplicación a proyectos para PyMEs, dentro de un marco definido por el usuario.

B. Métricas de Estudio

En el capítulo de la Solución, se han propuesto un conjunto de métricas de Datos, Modelos y Proyectos para el Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información definido por [85], que pueden tipificarse según la norma ISO/IEC 9126 [31] [32] [33] en: métricas básicas y métricas derivadas. Como se mencionó en el apartado II, sección *A* sub-sección *2*, las métricas básicas se obtienen directamente contando diferentes tipos de elementos que componen un proyecto de explotación de información, por ejemplo la cantidad de atributos o registros de una tabla, la cantidad de valores nulos o con errores en una tabla, atributo o registro, la cantidad de registros duplicados, el número de clases en un modelo, entre otros.

TABLA XVIII. MÉTRICAS PARA PROYECTOS – EVALUACIÓN DE RESULTADOS

Tarea: Evaluar los resultados		
Métrica	**Significado**	**Cálculo de la Métrica**
EPD (P)	Éxito de resultados del proceso de desarrollo del proyecto de explotación de información. Mide la bondad de los resultados de explotación de información obtenidos respecto a los criterios de éxito del proyecto. El valor de esta métrica es un indicador para de aceptación del proyecto. ▪ Si EPD (P) es BAJO indica que el proceso no responde a los criterios de éxito del proyecto. ▪ Si EPD (P) es MEDIO indica que las tareas del proceso deben ser revisadas y ajustadas. ▪ Si EPD (P) es ALTO indica que el proceso cumple con los criterios de éxito del proyecto. La unidad de medida de la métrica es *% (porcentaje)* y toma valores entre 0 y 1.	$EPD(P) = \frac{\sum NMI_i * Peso(i)}{NMOD}$ donde NMI_i es el número de modelos de explotación de información cuyo interés *i* de resultados es Alto, Medio o Bajo.
NMIA	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Alto. La unidad de medida de la métrica es *modelos* y toma valores entre 0 y 1.	
NMIB	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Bajo. La unidad de medida de la métrica es *modelos* y toma valores entre 0 y 1.	
NMIM	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Medio. La unidad de medida de la métrica es *modelos* y toma valores entre 0 y 1.	

TABLA XIX. MÉTRICAS PARA PROYECTOS – PRODUCCIÓN DEL INFORME FINAL

Tarea: Producir un reporte final		
Métrica	**Significado**	**Cálculo de la Métrica**
DEFZ (P)	Desvío del esfuerzo [63] para desarrollar el proyecto de explotación de información. Mide la variación entre el esfuerzo estimado al inicio del proyecto y el esfuerzo real del mismo, permite prever futuros costos, re-presupuestar y planificar el proyecto en caso de ser necesario. La unidad de medida de esta métrica es *% (porcentaje)* y toma valores mayores o menores a 0.	$DEFZ(P) = \frac{EFZE(P) - EFZR(P)}{EFZR(P)}$ donde EFZE (P) es el esfuerzo estimado al inicio del proyecto.
DRPY (P)	Duración real [63] del proyecto de explotación de información. Representa el tiempo total empleado por el grupo de trabajo para desarrollar el proyecto. Esta métrica se compara con el tiempo estimado al inicio del proyecto. La unidad de medida de esta métrica es *meses* y toma valores mayores a 0.	$DRPY(P) = \sum TRSubprc_i$ donde $TRSubprc_i$ es el tiempo real insumido para desarrollar las tareas de cada uno de los subprocesos Entendimiento del Negocio, Entendimiento de los Datos, Preparación de los Datos, Modelado, Evaluación y Entrega [84].
EFZR (P)	Esfuerzo real [63] aplicado para desarrollar el proyecto de explotación de información. Mide el trabajo necesario para realizar cada una de las tareas del proceso de desarrollo [84]. La unidad de medida de esta métrica es *persona-mes* y toma valores mayores a 0.	
PRGS (P)	Progreso del proyecto [63] de explotación de información. Mide el estado general del progreso del proyecto respecto de lo planificado. La unidad de medida de la métrica es en *días* y toma valores mayores o menores a 0.	$PRGS(P) = TESubprc_i - TRSubprc_i$ donde $TESubprc_i$ y $TRSubprc_i$ es el tiempo estimado y real insumido para desarrollar las tareas de cada uno de los subprocesos Entendimiento del Negocio, Entendimiento de los Datos, Preparación de los Datos, Modelado, Evaluación y Entrega [84].

Las métricas derivadas, son aquellas que se obtienen mediante fórmulas matemáticas, a partir de otras métricas básicas o derivadas.

Al resultar difícil obtener datos de proyectos de Explotación de Información reales para PyMEs, que permitan tener un conteo detallado (con valores concretos) de algunas métricas básicas, se decide estudiar el comportamiento de aquellas métricas de Datos (sección *C*), de Modelos (sección *D*) y de Proyectos (sección *E*), que pueden someterse a un proceso de simulación por el método Monte Carlo. El estudio experimental se realiza con el siguiente protocolo:

- *Paso 1*: Desarrollo de un banco de pruebas de N proyectos, donde se generan en una tabla los datos de proyectos de explotación de información con los valores de las variables experimentales independientes, dentro del marco de la clasificación de un proyecto de tamaño pequeño [65]. A cada proyecto generado se le aplican las métricas derivadas. Este paso del experimento es realizado mediante la utilización de una planilla de cálculo de Microsoft Excel, donde se registran en la misma todas las simulaciones realizadas, los valores de todas las variables independientes aplicadas (o sea, las métricas básicas) y de las variables dependientes (o sea, las métricas derivadas).
- *Paso 2*: Integrar estadísticamente la información obtenida generando los gráficos y tablas auxiliares que se consideren necesarios.
- *Paso 3*: Interpretar los resultados experimentales obtenidos y formular, como conclusión, una regla de comportamiento general de la métrica.

C. Estudio de las Métricas de Datos

Para estudiar el comportamiento de las Métricas de Datos, se decide generar un banco de pruebas simulado con diferentes cantidades de proyectos de explotación de información, considerando las restricciones indicadas en [65] para un proyecto de tamaño pequeño, al que se le aplican las métricas propuestas para el Entendimiento de los Datos (apartado IV, sección *B*, sub-sección *1.a*) y Preparación de los Datos (apartado IV, sección *B*, sub-sección *1.b*). La simulación de las métricas de Datos utiliza las siguientes variables independientes (sub-sección *1*) y dependientes (sub-sección *2*).

1. Variables Independientes

Las variables independientes que se van a generar mediante el proceso de simulación, son las correspondientes a las métricas básicas definidas en la sub-secciones de Entendimiento de Datos (apartado IV, sección *B*, sub-sección *1.a*) y Preparación de los Datos (apartado IV, sección *B*, sub-sección *1.b*) y que afectan directamente a las métricas derivadas. Para estas métricas básicas se define un valor específico o un valor aleatorio, considerando las restricciones por el tamaño del proyecto, restringiendo así la cantidad de combinaciones.

Las variables independientes a ser utilizadas se muestran en la Tabla XX.

TABLA XX. VARIABLES INDEPENDIENTES PARA MÉTRICAS DE DATOS

Variable Independiente (métrica básica)	Descripción
NA (T)	Número de atributos de la tabla, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NANC (T)	Número de atributos no correctos de la tabla, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NANS (T)	Número de atributos no significativos de la tabla, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NASE (T)	Número de atributos útiles sin errores en la tabla, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NAUD (T)	Número de atributos útiles pero con defectos en la tabla, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NR (T)	Número de registros de la tabla, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NT	Número de tablas, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NVE (T)	Número de valores erróneos en la tabla, considerando los valores erróneos por cada atributo o registro, cons un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NVN (T)	Número total de valores nulos o faltantes en la tabla, considerando los valores nulos o faltantes por cada atributo o registro, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.

2. Variables Dependientes

Para este proceso de simulación las variables dependientes, o sea las que son afectadas por las variables independientes, son los resultados de aplicar las fórmulas de las métricas derivadas de Densidad de Valores Nulos (sub-sección *3*), Nivel de Compleción (sub-sección *3*), Grado de Corrección de Datos de la Tabla (sub-sección *4*) y Grado de Utilidad de Atributos (sub-sección *5*), para las variables independientes definidas.

Las variables dependientes a ser utilizadas se muestran en la Tabla XXI.

TABLA XXI. VARIABLES DEPENDIENTES PARA MÉTRICAS DE DATOS

Variable Dependiente (métrica derivada)	Descripción
DVN (T)	Densidad de valores o nulos o faltantes en la tabla, la cual depende del número de valores nulos, número de registros y números de atributos que tengan las tablas.
GCD (T)	Grado de corrección de los datos de la tabla, la cual depende del número de valores nulos, número de valores erróneos, número de registros y números de atributos que tengan las tablas.
GUA (T)	Grado de utilidad de los atributos de las tablas para el proyecto, la cual depende del número de atributos útiles sin errores, útiles pero con defectos, no correctos y no significativos que tengan las tablas.
NCT (T)	Nivel de compleción de la tabla, la cual depende del resultado obtenido en la métrica DVN (T), ya que se interpreta como complemento de ésta.

La relación entre las variables independientes y dependientes para las Métricas de Datos, indicando como afectan unas a otras, puede verse en la Fig. 6.

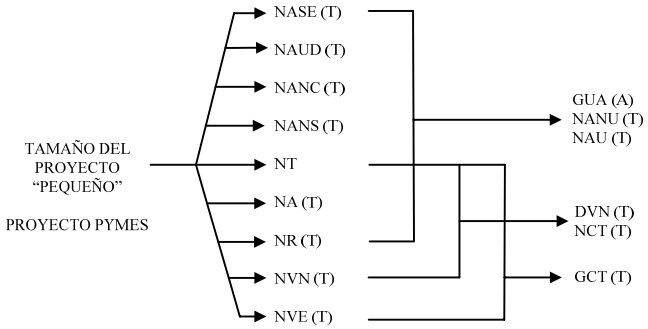

Fig. 6. Relación entre las Variables Independientes y Dependientes para Métricas de Datos

Cabe mencionar, que algunas de las métricas de datos sólo se calculan en función de la suma de los valores de otras métricas básicas y no se requieren estudiar su comportamiento. Estas métricas denominadas por la norma ISO/IEC 9126 como de agregación, se indican en la Tabla XXII.

3. Métrica de Densidad de Valores Nulos y Nivel de Compleción de la Tabla

Para estudiar las métricas de Densidad de Valores Nulos – DVN (T) y Nivel de Compleción de la Tabla – NVC (T), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de las mismas.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en las métricas DVN (T) y NCT (T), consideradas como métricas derivadas. La lista de variables experimentales se muestra en la Tabla XXIII.

En la sección de diseño experimental (sub-sección *b*) se establece como criterio para la simulación, los valores asociados a todas las tablas disponibles para el proyecto y no los valores por cada tabla específica. No obstante, los resultados experimentales y conclusiones obtenidas son de aplicación a cada tabla, atributo y registro en particular.

TABLA XXII. VARIABLES DEPENDIENTES PARA MÉTRICAS DE DATOS NO CONSIDERADAS EN EL ANÁLISIS

Variable Dependiente (métrica agregación)	Descripción
NA (TI)	Número de atributos que forman la tabla integrada, con los datos necesarios para construir los modelos para el proyecto de explotación de información, la cual se define como: $NA(TI) = NANI(TI) + \sum_{i} NAU(T_i)$ donde NANI (TI) son los atributos nuevos que se agregan a la tabla integrada. Mientras que el subíndice *i* representa aquellas tablas cuyo grado de utilidad de la métrica GUA (T) es MEDIO O ALTO.
NANU (T)	Número de atributos no útiles en la tabla, la cual se define como: $NANU(T) = NANS(T) + NANC(T)$
NAU (T)	Número de atributos útiles en la tabla, la cual se define como: $NAU(T) = NASE(T) + NAUD(T)$
NR (TI)	Número de registros que forman la tabla de datos integrada, con los casos para construir, entrenar y probar los modelos del proyecto de explotación de información, la cual se define como: $NR(TI) = \sum_{i} NRU(T_i)$ donde el subíndice *i* representa las tablas analizadas. Con estos registros se descubren los patrones de conocimiento útiles para el problema de negocio.
NRNU (T)	Número de registros no útiles en la tabla, la cual se define como: $NRNU(T) = NRNC(T) + NRDP(T)$ donde NRNC (T) y NRDP (T) es el número de registros no correctos y duplicados de cada tabla, respectivamente. Estos registros no se incluyen en la tabla de datos integrada a utilizar para construir los modelos del proyecto de explotación de información. Son eliminados en la tarea de Limpiar los Datos – subproceso Preparación de Datos.
NRU (T)	Número de registros útiles en la tabla, la cual se define como: $NRU(T) = NR(T) – NRNU(T)$ Estos registros forman parte de la tabla de datos integrada a utilizar para construir los modelos del proyecto de explotación de información.

b) Diseño Experimental

Para analizar el comportamiento de las métricas enunciadas, se utiliza un banco de pruebas simulado de 4.000 proyectos de explotación de información, definiendo un rango de valores específico para cada una de las variables experimentales independientes y considerando las restricciones indicadas en [65] según el tamaño del proyecto. De esta manera, se generan diferentes combinaciones sujetas a análisis. Los valores simulados para las variables experimentales independientes se muestran en la Tabla XXIV.

Para el caso de la variable NVN (T) indicada en la Tabla XXIV, los valores asignados se corresponden con los valores y porcentajes de valores nulos definidos para el modelo de estimación DMCoMo [48] [49], que se indican en la Tabla XXV. Por otra parte, cabe mencionar, que por tratarse de una métrica básica sometida a un proceso de simulación, no se puede asignar un número a priori de valores nulos que afectan a las tablas, atributos y registros utilizados en el proyecto. Por tal motivo, se decide establecer un rango de valores, con el fin de analizar el comportamiento general de las métricas DVN (T) y NCT (T), según este rango.

TABLA XXIII. VARIABLES EXPERIMENTALES PARA LAS MÉTRICAS DVN (T) Y NCT (T)

Variable Experimental	Descripción
DVN (T)	Densidad de valores nulos o faltantes en la tabla, la cual se define como: $DVN(T) = \frac{NVN(T)}{NR(T) * NA(T)}$
NCT (T)	Nivel de compleción de la tabla, la cual se define como: $NCT(T) = 1 – DVN(T)$
NA (T)	Número de atributos de la tabla. Si el número de atributos es elevado, implica un mayor análisis y preparación de los datos de la tabla, y en consecuencia un mayor esfuerzo para el proyecto de explotación de información.
NR (T)	Número de registros de la tabla. Si el número de atributos es elevado, implica un mayor análisis y preparación de los datos de la tabla, y en consecuencia un mayor esfuerzo para el proyecto de explotación de información.
NT	Número de tablas disponibles para el proyecto de explotación de información. Si el número de tablas es elevado, implica un mayor esfuerzo y tiempo de entendimiento, preparación y limpieza de los datos, para el proyecto de explotación de información.
NVN (T)	Número de valores nulos o faltantes en la tabla, considerando los valores nulos o faltantes por cada atributo o registro.

TABLA XXIV. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LAS MÉTRICAS DVN (T) Y NCT (T)

Variable Experimental	Descripción
NR (T)	Número de registros de la tabla, con un rango de valores específico 5.000.000, 10.000.000, 15.000.000, 20.000.000, 25.000.000, 30.000.000, 35.000.000, 40.000.000, 45.000.000 y 50.000.000, según lo predefinido para un proyecto de tamaño pequeño.
NA (T)	Número de atributos de la tabla, con un rango de valores específico 50, 100, 150, 200, 250, 300, 350, 400, 450 y 500, según lo predefinido para un proyecto de tamaño pequeño.
NT	Número de tablas, con un rango de valores específico 10, 20, 30, 40, 50, 60, 70 y 80, según lo predefinido para un proyecto de tamaño pequeño. Si bien, los valores de esta métrica básica no influyen directamente en el resultado de las métricas que se estudian, permite asociarla a la estimación del esfuerzo necesario [64; 47; 48] que requiere desarrollar un proyecto de explotación de información para PyMEs.
NVN (T)	Número de valores nulos o faltantes en la tabla, con un rango de valores específico 1, 2, 3, 4 y 5.

TABLA XXV. VALORES ASOCIADOS A LA MÉTRICA DVN (T)

Valor	Descripción
1	Hasta 10% de valores nulos
2	De 10 a 15% de valores nulos
3	De 15 a 20% de valores nulos
4	De 20 a 25% de valores nulos
5	Más de 25% de valores nulos

A partir de la asignación de los valores específicos a cada variable independiente, se generan los datos los proyectos de explotación de información. De esta manera se busca ver el comportamiento de las métricas DVN (T) y NCT (T) (variables

experimentales dependientes) de acuerdo a la clasificación definida en este trabajo para proyectos en PyMEs.

c) Ejecución y Resultado Experimental

En esta sección se estudia el comportamiento de las métricas DVN (T) y NCT (T), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que la métrica derivada DVN (T) mide la proporción de datos nulos que tiene/n la/s tabla/s en función del número total de datos disponibles (registros, atributos, valores nulos) para el proyecto, mientras que la métrica NCT (T) mide cuán completa está/n la/s tablas con datos. A partir de los datos obtenidos en los experimentos realizados, se observa que la densidad de valores nulos presenta una dispersión entre 0 (ausencia de valores nulos en las tablas) y 1 (todas las tablas con valores nulos), con una mediana de densidad de valores nulos de 0.17.

Por otra parte, se observa que la dispersión del nivel de compleción de las tablas también está entre 0 (todas las tablas con valores nulos) y 1 (ausencia de valores nulos en las tablas), con una mediana de nivel de compleción de 0.83. En el trabajo de [65], se menciona que en proyectos pequeños el porcentaje de valores nulos no debe superar el 15% de los datos totales, correspondiente a los valores 1 y 2 de la Tabla XXV, para que los mismos sean considerados de calidad para el desarrollo de un proyecto de explotación de información.

d) Regla Experimental

Como regla experimental del comportamiento de las métricas analizadas, se concluye que al aumentar el número de valores nulos en la/s tabla/s, la densidad de valores nulos tiende a su valor máximo (valor 1) y el nivel de compleción a su valor mínimo (valor 0). Lo contrario sucede si se disminuye el número de valores nulos.

En base a este análisis y a lo establecido en el trabajo de [65], se sugiere que el valor de la métrica DVN (T) sea inferior a 0.15 (o sea, 15% de los datos), mientras que la métrica NCT (T) sea superior a 0.85 (o sea, 85% de los datos), a fin de facilitar el análisis posterior de los datos y permitir una adecuada aplicación de los tipos de modelado para descubrimiento de conocimiento. Esta misma conclusión se aplica para analizar la densidad de valores nulos por cada atributo y registro de la/s tabla/s.

4. Métrica de Grado de Corrección de Datos de la Tabla

Para estudiar la métrica de Grado de Corrección de Datos de la Tabla – GCD (T), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de la misma.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en la métrica de estudio, la cual se toma como una métrica derivada. La lista de variables experimentales se muestra en la Tabla XXVI.

En la sección de diseño experimental (sub-sección *b*) se establece como criterio para la simulación, los valores asociados a todas las tablas de un proyecto y no los valores por cada tabla específica. No obstante, los resultados experimentales y conclusiones obtenidas son de aplicación a cada tabla, atributo y registro en particular.

b) Diseño Experimental

Para analizar el comportamiento de la métrica GCD (T), se utiliza un banco de pruebas simulado de 8.000 proyectos de explotación de información, definiendo un rango de valores específico para cada una de las variables experimentales independientes y considerando las restricciones indicadas en [65] según el tamaño del proyecto. De esta manera, se generan diferentes combinaciones sujetas a análisis. Los valores simulados para las variables experimentales independientes se muestran en la Tabla XXVII.

TABLA XXVI. VARIABLES EXPERIMENTALES PARA LA MÉTRICA GCD (T)

Variable Experimental	Descripción
GCD (T)	Grado de corrección de los datos de la tabla, la cual se define como: $GCD(T) = 1 - \frac{NVE(T) + NVN(T)}{NR(T) * NA(T)}$
NA (T)	Número de atributos de la tabla. Si el número de atributos es elevado, implica un mayor análisis y preparación de los datos de la tabla, y en consecuencia un mayor esfuerzo para el proyecto de explotación de información.
NR (T)	Número de registros de la tabla. Si el número de atributos es elevado, implica un mayor análisis y preparación de los datos de la tabla, y en consecuencia un mayor esfuerzo para el proyecto de explotación de información.
NT	Número de tablas disponibles para el proyecto de explotación de información. Si el número de tablas es elevado, implica un mayor esfuerzo y tiempo de entendimiento, preparación y limpieza de los datos, para el proyecto de explotación de información.
NVE (T)	Número de valores erróneos en la tabla, considerando los valores erróneos por cada atributo o registro.
NVN (T)	Número de valores nulos o faltantes en la tabla, considerando los valores nulos o faltantes por cada atributo o registro.

TABLA XXVII. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LA MÉTRICA GCD (T)

Variable Experimental	Descripción
NA (T)	Número total de atributos de las tablas, con un rango de valores específico 50, 100, 150, 200, 250, 300, 350, 400, 450 y 500, según lo predefinido para un proyecto de tamaño pequeño.
NR (T)	Número total de registros de las tablas, con un rango de valores específico 5.000.000, 10.000.000, 15.000.000, 20.000.000, 25.000.000, 30.000.000, 35.000.000, 40.000.000, 45.000.000 y 50.000.000, según lo predefinido para un proyecto de tamaño pequeño.
NT	Número de tablas, con un rango de valores específico 10, 20, 30, 40, 50, 60, 70 y 80, según lo predefinido para un proyecto de tamaño pequeño. Si bien, los valores de esta métrica básica no influyen directamente en el resultado de las métricas que se estudian, permite asociarla a la estimación del esfuerzo necesario [48] [49] [65] que requiere desarrollar un proyecto de explotación de información para PyMEs.
NVN (T)	Número de valores nulos o faltantes en las tablas, con un rango de valores específico 1, 2.

Cabe mencionar que para este análisis, se establece como rango para la métrica NVN (T), indicada en la Tabla XXVII, los valores 1 y 2 de la Tabla XXV, que se corresponden con las conclusiones sugeridas para la métrica de densidad de valores nulos DVN (T), de la sub-sección *3.d*.

A su vez, en el caso de los valores definidos para la métrica NVE (T), los mismos se establecen siguiendo el criterio que se indicó para la métrica NVN (T) en la sub-sección *3.b*, y que se indican en la Tabla XXVIII.

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de la métrica GCD (T), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que la métrica GCD (T) siempre toma un valor ALTO, considerándose como parámetro para este análisis un valor superior a 0.70, cuando el rango de valores de la métrica NVE (T) es 1 ó 2, es decir que el número de valores erróneos es inferior al 15% de los datos totales de la/s tabla/s y el número de valores nulos también se ubica por debajo de este porcentaje. En la Fig. 7 se visualiza el comportamiento de la métrica GCD (T), en función del rango de valores nulos y de valores erróneos en los datos de la/s tabla/s.

TABLA XXVIII. VALORES ASOCIADOS A LA MÉTRICA NVE (T)

Valor	Descripción
1	Hasta 10% de valores erróneos
2	De 10 a 15% de valores erróneos
3	De 15 a 20% de valores erróneos
4	De 20 a 25% de valores erróneos
5	Más de 25% de valores erróneos

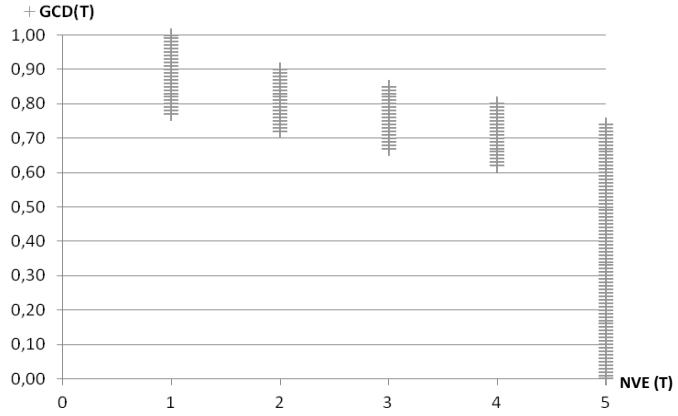

Fig. 7. Comportamiento de la métrica GCD (T) cuando NVN (T) es 1 ó 2

Cuando el rango de la métrica NVE (T) toma un valor 3, se observa que la mayoría de los proyectos analizados toman un valor ALTO como métrica del grado de corrección de la/s tabla/s, pero otros proyectos toman un valor levemente inferior, dependiendo fundamentalmente del número de valores nulos que tenga/n la/s tabla/s. Cuando el rango de la métrica NVE (T) toma un valor 4, se observa que la mitad de los proyectos analizados toman un valor ALTO para la métrica GCD (T), mientras que la otra mitad de los proyectos toma un valor por debajo de 0.70, considerándose BAJO. Cuando el rango de la métrica NVE (T) toma un valor 5, en prácticamente todos los proyectos analizados se observa que el grado de corrección de la/s tabla/s es BAJO, dejando a las mismas inconsistentes para el proyecto.

d) Regla Experimental

Como regla experimental del comportamiento de la métrica analizada, se concluye que al aumentar el número de valores erróneos en la/s tabla/s, con variaciones del número de valores nulos menores al 15% de los datos, el grado de corrección de los datos la tabla GCD (T) tiende a ser BAJO. Al igual que con el número de valores nulos, se sugiere que el número de valores erróneos se ubique por debajo del 15% del total de datos disponibles, para lograr una métrica GCD (T) con ALTO grado de corrección de la/s tabla/s. La misma conclusión aplica por cada atributo y registro de la/s tabla/s, permitiendo identificar los atributos y registros con más valores erróneos. No obstante, el valor utilizado como parámetro para esta métrica dependerá del criterio que adopte el grupo de desarrollo del proyecto y del número de valores nulos y erróneos que se definan como límites, en función de la cantidad de tablas y datos disponibles.

5. Métrica de Grado de Utilidad de Atributos

Para estudiar la métrica de Grado de Utilidad de Atributos de la tabla – GUA (T), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de la misma.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en la métrica de estudio, la cual se toma como una métrica derivada. La lista de variables experimentales se muestra en la Tabla XXIX.

TABLA XXIX. VARIABLES EXPERIMENTALES PARA LA MÉTRICA GUA (T)

Variable Experimental	Descripción
GUA (T)	Grado de utilidad de los atributos de la tabla, la cual se define como: $GUA(T) = \frac{NA(T) - (NO\_UTILES(T) + 0{,}5 * NAUD(T))}{NA(T)}$ donde $NO\_UTILES(T) = NANC(T) + NANS(T)$ son los atributos no útiles para el proyecto de explotación de información.
NA (T)	Número de atributos de la tabla. Si el número de atributos es elevado, implica un mayor análisis y preparación de los datos de la tabla, y en consecuencia un mayor esfuerzo para el proyecto de explotación de información.
NANC (T)	Número de atributos no correctos de la tabla.
NANS (T)	Número de atributos no significativos de la tabla.
NASE (T)	Número de atributos útiles si errores en la tabla.
NAUD (T)	Número de atributos útiles pero con defectos en la tabla.
NT	Número de tablas disponibles para el proyecto de explotación de información. Si el número de tablas es elevado, implica un mayor esfuerzo y tiempo de entendimiento, preparación y limpieza de los datos, para el proyecto de explotación de información.

En la sección de diseño experimental (sub-sección *b*) se establece como criterio para la simulación, los valores asociados a todas las tablas de un proyecto y no los valores por cada tabla específica. No obstante, los resultados experimentales y conclusiones obtenidas son de aplicación a cada tabla en particular.

b) Diseño Experimental

Para analizar el comportamiento de la métrica GUA (T), se utiliza un banco de pruebas simulado de 22.880 proyectos de explotación de información, definiendo un rango de valores específico para cada una de las variables experimentales independientes y considerando las restricciones indicadas en [64] según el tamaño del proyecto. De esta manera, se generan diferentes combinaciones sujetas a análisis. Los valores simulados para las variables experimentales independientes se muestran en la Tabla XXX.

El resultado y comportamiento de la métrica derivada GUA (T) está asociado al número de atributos útiles sin errores, útiles con defectos, no correctos (con muchos errores y valores nulos) y no significativos que pueden presentar las tablas en un proyecto de explotación de información, cuyas métricas básicas

NASE (T), NAUD (T), NANC (T) y NANS (T), se indican en la Tabla XXX.

A su vez, las primeras dos métricas representan la cantidad total de atributos útiles para el desarrollo del proyecto, mientras que las dos últimas métricas representan la cantidad total de atributos no útiles.

TABLA XXX. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LA MÉTRICA GUA (T)

Variable Experimental	Descripción
NA (T)	Número de atributos de la tabla, con un rango de valores específico 50, 100, 150, 200, 250, 300, 350, 400, 450 y 500, según lo predefinido para un proyecto de tamaño pequeño.
NANC (T)	Número de atributos no correctos de la tabla, con un rango de valores específico 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 y 100. Este rango de valores se expresa como un porcentaje del total de atributos de la/s tabla/s.
NANS (T)	Número de atributos no significativos de la tabla, con un rango de valores específico 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 y 100. Este rango de valores se expresa como un porcentaje del total de atributos de la/s tabla/s.
NASE (T)	Número de atributos útiles sin errores en la tabla, con un rango de valores específico 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 y 100. Este rango de valores se expresa como un porcentaje del total de atributos de la/s tabla/s.

Por otra parte, se tiene como restricción que la suma de estas cuatro métricas deben cubrir la totalidad de los atributos de la/s tabla/s.

$$NA(T) = NAU(T) + NANU(T)$$

donde:

- $NAU(T) = NASE(T) + NAUD(T)$ son los atributos útiles para el proyecto.
- $NANU(T) = NANC(T) + NANS(T)$ son los atributos no útiles para el proyecto.

Al tratarse de métricas básicas sometidas a un proceso de simulación, no se les puede asignar un número a priori de atributos que afectan a las tablas. Por tal motivo, se decide establecer un rango de valores porcentuales con intervalos de a 10% (entre 0 y 100%), respecto del número total de atributos, con el fin de analizar el comportamiento general de la métrica GUA (T), según estos rangos. Por otra parte, y como se indicó anteriormente, existe una restricción entre estas cuatro métricas respecto del total de atributos de la/s tabla/s. Por consiguiente, se necesitan sólo aquellas configuraciones de las métricas NASE (T), NAUD (T), NANC (T) y NANS (T), cuya suma represente el 100% de atributos. De este análisis se obtienen 286 configuraciones posibles aplicables a un proyecto.

La selección de atributos es una actividad muy importante del pre-procesamiento de datos cuando se desea realizar descubrimiento de conocimiento en bases de datos. Su principal objetivo es eliminar atributos no útiles para obtener problemas computacionalmente tratables, sin afectar la calidad de la solución [26].

Al inicio de todo proyecto de explotación de información, en particular en la etapa de preparación de los datos, es deseable disponer de la mayor cantidad de atributos útiles que permitan la selección de aquellos que sean necesarios para la generación de los modelos del proyecto. Del mismo modo, se busca tener la menor cantidad posible de atributos no útiles en las tablas a utilizar, con el fin de favorecer la viabilidad en la realización del proyecto. En el trabajo de [65] se indica que en proyectos de tamaño pequeño, los atributos no poseen gran variedad de valores erróneos ni nulos, sino que los mismos son de buena calidad.

Sin embargo, no existe un criterio universal que defina qué indicadores se deben considerar como atributos útiles y no útiles en un proyecto de explotación de información. Por consiguiente, y a partir de la propia experiencia, en las Tablas XXXI y XXXII se indican los valores y rangos contemplados para un proyecto de explotación de información para PyMEs, de acuerdo a la clasificación de atributos útiles y no útiles. Además, se enumeran los valores asociados y utilizados para la simulación.

TABLA XXXI. RANGOS Y VALORES RELACIONADOS A LAS MÉTRICAS NASE (T) Y NAUD (T) PARA ATRIBUTOS ÚTILES

Rango	Descripción	Valores de simulación
Muy Pocos	De 0 a 20% de atributos	0, 10, 20
Pocos	De 21 a 30% de atributos	30
Normal	De 31 a 60% de atributos	40, 50, 60
Bastantes	De 61 a 70% de atributos	70
Muchos	Más de 70% de atributos	80, 90, 100

TABLA XXXII. RANGOS Y VALORES RELACIONADOS A LAS MÉTRICAS NANC (T) Y NANS (T) PARA ATRIBUTOS NO ÚTILES

Rango	Descripción	Valores de simulación
Muy Pocos	De 0 a 10% de atributos	0, 10
Pocos	De 11 a 25% de atributos	20
Normal	De 26 a 40% de atributos	30, 40
Bastantes	De 41 a 60% de atributos	50, 60
Muchos	Más de 60% de atributos	70, 80, 90, 100

A partir de la asignación de los valores a cada variable independiente, se generan los datos de los proyectos de explotación de información. De esta manera se busca obtener los resultados y el comportamiento de la métrica GUA (T) (variable experimental dependiente) en función de la variación de estas variables. En la Tabla XXXIII se indican los rangos y valores que se consideran como parámetros para la métrica en estudio.

TABLA XXXIII. RANGOS Y VALORES RELACIONADOS A LA MÉTRICA GUA (T)

Rango	Descripción
Bajo	De 0 a 30% de utilidad de atributos
Medio	De 31 a 70% de utilidad de atributos
Alto	Más de 70% de utilidad de atributos

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de la métrica GUA (T), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que la métrica derivada GUA (T) toma un valor ALTO, considerándose como parámetro para este análisis un valor relativo superior al 70%, cuando el porcentaje de atributos útiles sin errores, asociado a la métrica NASE (T), está entre 50 y 100% del total de atributos de la/s tabla/s. A su vez, el porcentaje de atributos útiles con defectos, asociado a la métrica NAUD (T), se ubica

entre 0 y 50% de los atributos totales. En la Tabla XXXIV, se muestra la distribución del número de proyectos en función del porcentaje de atributos útiles sin errores y útiles con defectos. En esta tabla, se observa que para que el grado de utilidad de los atributos sea ALTO se necesita entre el 80 y 100% de atributos útiles para el proyecto. Además, como mínimo es necesario un 50% de atributos útiles sin errores y un 50% de atributos útiles con defectos (ambos categorizados como rango "normal"), lo cual implica tener el 100% de atributos totales útiles y ningún tipo de atributo no útil.

TABLA XXXIV. DISTRIBUCIÓN DE PROYECTOS CON VALOR GUA (T)=ALTO RESPECTO DE LAS MÉTRICAS NASE (T) Y NAUD (T)

		% Útiles sin Errores – NASE (T)						
		50	*60*	*70*	*80*	*90*	*100*	*Total*
% Útiles con Defectos NAUD (T)	***0***	0	0	0	240	160	80	**480**
	10	0	0	240	160	80	0	**480**
	20	0	0	160	80	0	0	**240**
	30	0	160	80	0	0	0	**240**
	40	0	80	0	0	0	0	**80**
	50	80	0	0	0	0	0	**80**
	Total	**80**	**240**	**480**	**480**	**240**	**80**	***1600***

Por otra parte, el porcentaje de atributos no correctos y no significativos, asociados a las métricas NANC (T) y NANS (T), está entre 0 y 20% del total de atributos de la/s tabla/s. En la Tabla XXXV, se muestra la distribución del número de proyectos en función del porcentaje de atributos no correctos y no significativos. En la misma, se observa que para que el grado de utilidad de los atributos sea ALTO se necesita un máximo de 20% de atributos no útiles para el proyecto.

Con esta información, se decide estudiar el comportamiento de aquellos proyectos que mejor caracterizan el grado de la métrica GUA (T) con valor ALTO.

TABLA XXXV. DISTRIBUCIÓN DE PROYECTOS CON VALOR GUA (T)=ALTO RESPECTO DE LAS MÉTRICAS NANC (T) Y NANS (T)

		% No Correctos NANC (T)			
		0	*10*	*20*	*Total*
% No Significativos NANS (T)	***0***	480	320	160	**960**
	10	320	160	0	**480**
	20	160	0	0	**160**
	Total	**960**	**480**	**160**	***1600***

Del análisis experimental, se observa que la métrica GUA (T) instancia una mayor cantidad de proyectos con valor ALTO, cuando el porcentaje de atributos útiles sin errores, asociado a la métrica NASE (T), está entre 70 y 80% del total de atributos de la/s tabla/s. La variación de los rangos de esta métrica respecto al total de proyectos simulados se muestra en la Fig. 8. A su vez, el porcentaje de atributos útiles con defectos, asociado a la métrica NAUD (T), se ubica entre 0 y 20% de los atributos totales. Esta proporción de atributos útiles representa el 55% de los proyectos simulados con ALTO grado de utilidad de atributos, obteniéndose un valor medio relativo para la métrica GUA (T) del 80%. En la Tabla XXXVI, se indican aquellos proyectos con mayor incidencia en la métrica GUA (T) en función del porcentaje de atributos útiles sin errores y útiles con defectos.

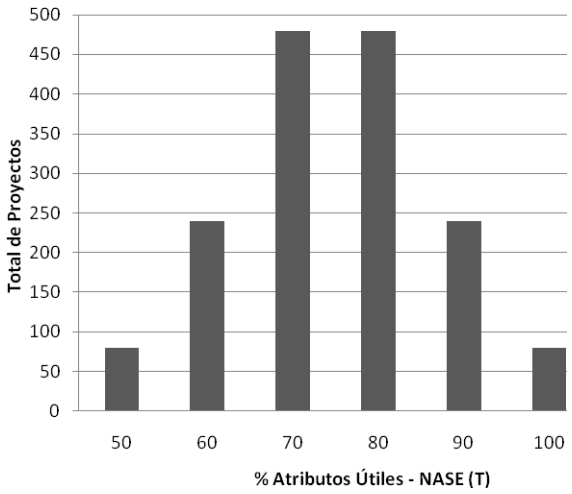

Fig. 8. Variación de rangos de la métrica NASE (T) con GUA (T)=Alto

TABLA XXXVI. PROYECTOS CON MAYOR INCIDENCIA SOBRE GUA (T)=ALTO RESPECTO DE LAS MÉTRICAS NASE (T) Y NAUD (T)

		% Útiles sin Errores – NASE (T)						
		50	*60*	*70*	*80*	*90*	*100*	*Total*
% Útiles con Defectos NAUD (T)	***0***	0	0	**0**	**240**	160	80	**480**
	10	0	0	**240**	**160**	80	0	**480**
	20	0	0	**160**	**80**	0	0	**240**
	30	0	160	80	0	0	0	**240**
	40	0	80	0	0	0	0	**80**
	50	80	0	0	0	0	0	**80**
	Total	**80**	**240**	**480**	**480**	**240**	**80**	***1600***

Por otra parte, para los proyectos que más inciden sobre la métrica GUA (T), se observa que en mayor proporción el porcentaje de atributos no correctos y no significativos, asociados a las métricas NANC (T) y NANS (T), se ubica entre 0 y 10% del total de atributos de la/s tabla/s, tal como se muestra en la Tabla XXXVII. Esta proporción de atributos no útiles representa el 64% de los proyectos con mayor incidencia sobre la métrica en estudio, mientras que el 36% restante presenta un comportamiento extrapolado respecto a los atributos no correctos y no significativos.

TABLA XXXVII. PROYECTOS CON MAYOR INCIDENCIA SOBRE GUA (T)=ALTO RESPECTO DE LAS MÉTRICAS NANC (T) Y NANS (T)

		% No Correctos NANC (T)			
		0	*10*	*20*	*Total*
% No Significativos NANS (T)	***0***	**80**	**160**	160	**400**
	10	**160**	**160**	0	**320**
	20	160	0	0	**160**
	Total	**400**	**320**	**160**	***880***

Del análisis realizado a los proyectos que más inciden cuando la métrica GUA (T) toma un valor ALTO, se concluye que cuando el número de atributos útiles sin errores está entre 70 y 80%, el número de atributos útiles con defectos está entre 0 y 20% y el número de atributos no correctos y no significativos está entre 0 y 10%, se obtiene la mayor cantidad

de proyectos con características similares y un valor ALTO para la métrica GUA (T). Esto representa el 35% de los proyectos simulados con ALTO grado de utilidad de atributos, obteniéndose un valor medio relativo para la métrica GUA (T) del 82%. Sin embargo, cuando el número de atributos no correctos o no significativos toma el valor 20% (una de las dos métricas toma valor 0%) manteniendo el número de atributos útiles, el valor medio relativo para la métrica GUA (T) desciende al 77%, lo que representa un 20% de los proyectos simulados.

En un segundo análisis experimental, se observa que la métrica derivada GUA (T) toma un valor MEDIO, considerándose como parámetro para este análisis un valor relativo entre 31 y 70%, cuando el porcentaje de atributos útiles sin errores, asociado a la métrica NASE (T), está entre 0 y 70% del total de atributos de la/s tabla/s. A su vez, el porcentaje de atributos útiles con defectos, asociado a la métrica NAUD (T), se ubica entre 0 y 100% de los atributos totales. En la Tabla XXXVIII, se muestra la distribución del número de proyectos en función del porcentaje de atributos útiles sin errores y útiles con defectos. En la misma, se observa que la mayoría de los proyectos tienen una proporción de atributos útiles que está entre 40 y 80% del total de los atributos de la/s tabla/s, lo cual puede considerarse como el indicador cuando el grado de utilidad de los atributos es MEDIO.

Obsérvese que los casos cuya cantidad de atributos útiles supera el 80% podrían entenderse como de ALTO grado de utilidad para la métrica GUA (T). Sin embargo, la proporción de atributos útiles sin errores es menor cuando la métrica GUA (T) toma un valor MEDIO que cuando adquiere un valor ALTO. Lo contrario ocurre con la proporción de atributos útiles sin defectos, donde la misma es mayor cuando la métrica GUA (T) toma un valor MEDIO, implicando un mayor esfuerzo en la adecuación y normalización de los datos del atributo para ser utilizados en el proyecto.

TABLA XXXVIII. DISTRIBUCIÓN DE PROYECTOS CON VALOR GUA (T)=MEDIO RESPECTO DE LAS MÉTRICAS NASE (T) Y NAUD (T)

		% Útiles sin Errores - NASE (T)								
		0	***10***	***20***	***30***	***40***	***50***	***60***	***70***	***Total***
% Útiles con Defectos NAUD (T)	***0***	0	0	0	0	560	480	400	320	**1760**
	10	0	0	0	560	480	400	320	0	**1760**
	20	0	0	0	480	400	320	240	0	**1440**
	30	0	0	480	400	320	240	0	0	**1440**
	40	0	0	400	320	240	160	0	0	**1120**
	50	0	400	320	240	160	0	0	0	**1120**
	60	0	320	240	160	80	0	0	0	**800**
	70	320	240	160	80	0	0	0	0	**800**
	80	240	160	80	0	0	0	0	0	**480**
	90	160	80	0	0	0	0	0	0	**240**
	100	80	0	0	0	0	0	0	0	**80**
	Total	**800**	**1200**	**1680**	**2240**	**2240**	**1600**	**960**	**320**	***11040***

Por otra parte, el porcentaje de atributos no correctos y no significativos, asociados a las métricas NANC (T) y NANS (T), es menor al 60% del total de atributos de la/s tabla/s. En la Tabla XXXIX, se muestra la distribución del número de proyectos en función del porcentaje de atributos no correctos y no significativos. En la misma, se observa que para que el grado de utilidad de los atributos sea MEDIO se necesita un máximo de 60% de atributos no útiles para el proyecto.

TABLA XXXIX. DISTRIBUCIÓN DE PROYECTOS CON VALOR GUA (T)=MEDIO RESPECTO DE LAS MÉTRICAS NANC (T) Y NANS (T)

		% No Correctos - NANC (T)							
		0	***10***	***20***	***30***	***40***	***50***	***60***	***Total***
% No Significativos NANS (T)	***0***	400	480	560	640	480	320	160	**3040**
	10	480	560	640	480	320	160	0	**2640**
	20	560	640	480	320	160	0	0	**2160**
	30	640	480	320	160	0	0	0	**1600**
	40	480	320	160	0	0	0	0	**960**
	50	320	160	0	0	0	0	0	**480**
	60	160	0	0	0	0	0	0	**160**
	Total	**3040**	**2640**	**2160**	**1600**	**960**	**480**	**160**	***11040***

Con esta información, se decide estudiar el comportamiento de aquellos proyectos que mejor caracterizan el grado de la métrica GUA (T) con valor MEDIO.

Del análisis experimental, se observa que la métrica GUA (T) instancia una mayor cantidad de proyectos con valor MEDIO, cuando el porcentaje de atributos útiles sin errores, asociado a la métrica NASE (T), está entre 30 y 60% del total de atributos de la/s tabla/s. La variación de los rangos de esta métrica respecto al total de proyectos simulados se muestra en la Fig 9. A su vez, el porcentaje de atributos útiles con defectos, asociado a la métrica NAUD (T), se ubica entre 0 y 40% de los atributos totales. Esta proporción de atributos útiles representa el 57% de los proyectos simulados con grado de utilidad de atributos MEDIO, obteniéndose un valor medio relativo para la métrica GUA (T) del 51%. En la Tabla XL, se indican aquellos proyectos con mayor incidencia en la métrica GUA (T) en función del porcentaje de atributos útiles sin errores y útiles con defectos.

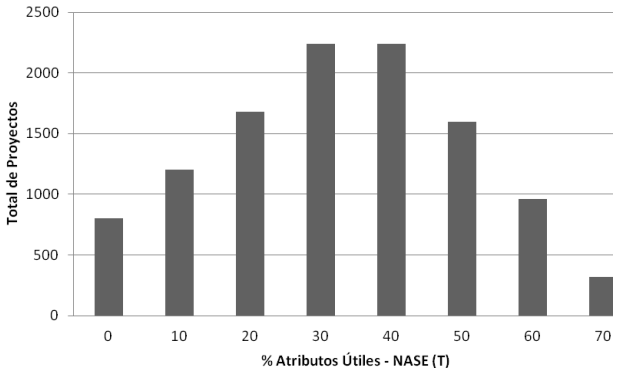

Fig. 9. Variación de rangos de la métrica NASE (T) con GUA (T)=Medio

Por otra parte, para los proyectos que más inciden sobre la métrica GUA (T), se observa que en mayor proporción el porcentaje de atributos no correctos y no significativos, asociados a las métricas NANC (T) y NANS (T), se ubica entre 0 y 30% del total de atributos de la/s tabla/s, tal como se muestra en la Tabla XLI. Esta proporción de atributos no útiles representa el 60% de los proyectos con mayor incidencia sobre la métrica en estudio, mientras que el 40% restante presenta un comportamiento extrapolado respecto a los atributos no correctos y no significativos.

TABLA XL. PROYECTOS CON MAYOR INCIDENCIA SOBRE GUA (T)=MEDIO RESPECTO DE LAS MÉTRICAS NASE (T) Y NAUD (T)

		% Útiles sin Errores - NASE (T)								
		0	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*Total*
% Útiles con Defectos NAUD (T)	*0*	0	0	0	**0**	**560**	**480**	**400**	320	**1760**
	10	0	0	0	**560**	**480**	**400**	**320**	0	**1760**
	20	0	0	0	**480**	**400**	**320**	**240**	0	**1440**
	30	0	0	480	**400**	**320**	**240**	**0**	0	**1440**
	40	0	0	400	**320**	**240**	**160**	**0**	0	**1120**
	50	0	400	320	240	160	0	0	0	**1120**
	60	0	320	240	160	80	0	0	0	**800**
	70	320	240	160	80	0	0	0	0	**800**
	80	240	160	80	0	0	0	0	0	**480**
	90	160	80	0	0	0	0	0	0	**240**
	100	80	0	0	0	0	0	0	0	**80**
	Total	**800**	**1200**	**1680**	**2240**	**2240**	**1600**	**960**	**320**	***11040***

Del análisis realizado a los proyectos que más inciden cuando la métrica GUA (T) toma un valor MEDIO, se concluye que cuando el número de atributos útiles sin errores está entre 30 y 60%, el número de atributos útiles con defectos está entre 0 y 40 % y el número de atributos no correctos y no significativos está entre 0 y 30%, se obtiene la mayor cantidad de proyectos con características similares y un valor MEDIO para la métrica GUA (T).

Esto representa el 34% de los proyectos simulados con un grado MEDIO de utilidad de atributos, obteniéndose un valor medio relativo para la métrica GUA (T) del 56%. Sin embargo, cuando el número de atributos no correctos o no significativos aumenta a valores entre 40 y 60% (una de las dos métricas decrece a valores entre 0 y 20%) manteniendo el número de atributos útiles, el valor medio relativo para la métrica GUA (T) desciende al 44%, lo que representa un 23% de los proyectos simulados.

TABLA XLI. PROYECTOS CON MAYOR INCIDENCIA SOBRE GUA (T)=MEDIO RESPECTO DE LAS MÉTRICAS NANC (T) Y NANS (T)

		% No Correctos - NANC (T)							
		0	*10*	*20*	*30*	*40*	*50*	*60*	*Total*
% No Significativos NANS (T)	*0*	**0**	**80**	**240**	**320**	320	240	160	**1360**
	10	**80**	**240**	**320**	**320**	240	160	0	**1360**
	20	**240**	**320**	**320**	**240**	160	0	0	**1280**
	30	**320**	**320**	**240**	**160**	0	0	0	**1040**
	40	320	240	160	0	0	0	0	**720**
	50	240	160	0	0	0	0	0	**400**
	60	160	0	0	0	0	0	0	**160**
	Total	**1360**	**1360**	**1280**	**1040**	**720**	**400**	**160**	***6320***

En el último análisis experimental realizado sobre la métrica GUA (T), se observa que la misma toma un valor BAJO, considerándose como parámetro para este análisis un valor relativo hasta 30%, cuando el porcentaje de atributos útiles sin errores, asociado a la métrica NASE (T), está entre 0 y 30% del total de atributos de la/s tabla/s. A su vez, el porcentaje de atributos útiles con defectos, asociado a la métrica NAUD (T), se ubica entre 0 y 60% de los atributos totales. En la Tabla XLII, se muestra la distribución del número de proyectos en función del porcentaje de atributos útiles sin errores y útiles con defectos. No obstante, se observa que la mayoría de los proyectos tienen una cantidad de atributos útiles que no supera el 40% del total de los atributos de la/s tabla/s, lo cual puede suponerse como un indicador cuando el grado de utilidad de los atributos es BAJO.

TABLA XLII. DISTRIBUCIÓN DE PROYECTOS CON VALOR GUA (T)=BAJO RESPECTO DE LAS MÉTRICAS NASE (T) Y NAUD (T)

		% Útiles sin Errores - NASE (T)				
		0	*10*	*20*	*30*	*Total*
% Útiles con Defectos NAUD (T)	*0*	880	800	720	640	**3040**
	10	800	720	640	0	**2160**
	20	720	640	560	0	**1920**
	30	640	560	0	0	**1200**
	40	560	480	0	0	**1040**
	50	480	0	0	0	**480**
	60	400	0	0	0	**400**
	Total	**4480**	**3200**	**1920**	**640**	***10240***

Por otra parte, el porcentaje de atributos no correctos y no significativos, asociados a las métricas NANC (T) y NANS (T), está entre 0 y 100% del total de atributos de la/s tabla/s. En la Tabla XLIII, se muestra la distribución del número de proyectos en función del porcentaje de atributos no correctos y no significativos. En la misma, se observa que para que el grado de utilidad de los atributos sea BAJO se necesita un mínimo de 40% de atributos no útiles para el proyecto.

Con esta información, se decide estudiar el comportamiento de aquellos proyectos que mejor caracterizan el grado de la métrica GUA (T) con valor BAJO.

Del análisis experimental, se observa que la métrica GUA (T) instancia una mayor cantidad de proyectos con valor BAJO, cuando el porcentaje de atributos útiles sin errores, asociado a la métrica NASE (T), no supera el 10% del total de atributos de la/s tabla/s. La variación de los rangos de esta métrica respecto al total de proyectos simulados se muestra en la Fig. 10. Con este parámetro, el porcentaje de atributos útiles con defectos, asociado a la métrica NAUD (T), no supera el 40% de los atributos totales. Esta proporción de atributos útiles representa el 66% de los proyectos simulados, con un valor medio relativo para la métrica GUA (T) del 14%.

Sin embargo, se observa también que si el porcentaje de atributos útiles asociado a las métricas NASE (T) y NAUD (T) no superan el 20% del total de atributos de la/s tabla/s, se representa una cantidad levemente inferior de proyectos (63%) pero con el mismo valor medio relativo para la métrica GUA (T).

Si bien en ambos casos se obtienen conclusiones similares, se puede considerar que con la segunda alternativa se consigue una distribución más pareja de atributos útiles sin errores y útiles con defectos, no superando entre ambos tipos de atributos útiles el 40% del total de la/s tabla/s. En la Tabla XLIV, se indican, para ambos casos, aquellos proyectos con mayor incidencia en la métrica GUA (T) en función del porcentaje de atributos útiles sin errores y útiles con defectos.

TABLA XLIII. DISTRIBUCIÓN DE PROYECTOS CON VALOR GUA (T)=BAJO RESPECTO DE LAS MÉTRICAS NANC (T) Y NANS (T)

		% No Correctos - NANC (T)											
		0	***10***	***20***	***30***	***40***	***50***	***60***	***70***	***80***	***90***	***100***	***Total***
% No Significativos NANS (T)	***0***	0	0	0	0	80	160	240	320	240	160	80	**1280**
	10	0	0	0	80	160	240	320	240	160	80	0	**1280**
	20	0	0	80	160	240	320	240	160	80	0	0	**1280**
	30	0	80	160	240	320	240	160	80	0	0	0	**1280**
	40	80	160	240	320	240	160	80	0	0	0	0	**1280**
	50	160	240	320	240	160	80	0	0	0	0	0	**1200**
	60	240	320	240	160	80	0	0	0	0	0	0	**1040**
	70	320	240	160	80	0	0	0	0	0	0	0	**800**
	80	240	160	80	0	0	0	0	0	0	0	0	**480**
	90	160	80	0	0	0	0	0	0	0	0	0	**240**
	100	80	0	0	0	0	0	0	0	0	0	0	**80**
	Total	**1280**	**1280**	**1280**	**1280**	**1280**	**1200**	**1040**	**800**	**480**	**240**	**80**	***10240***

Por otra parte, en la segunda alternativa de proyectos que más inciden sobre la métrica GUA (T), se observa una distribución simétrica de proyectos cuando el número de atributos no correctos y no significativos, asociados a las métricas NANC (T) y NANS (T), oscila entre 60 y 100% del total de atributos de la/s tabla/s. En este caso, todos los proyectos mantienen la misma relación para los atributos no útiles, tal como se muestra en la Tabla XLV.

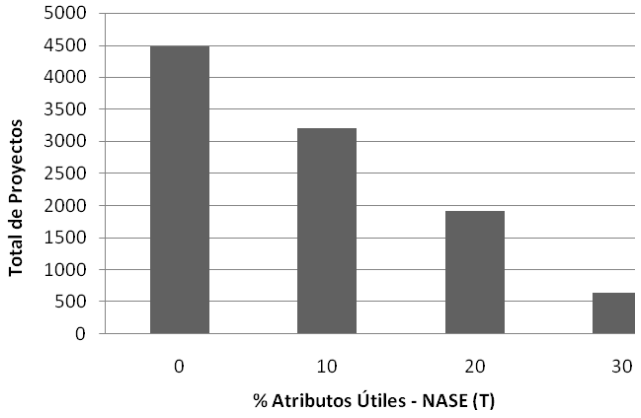

Fig. 10. Variación de rangos de la métrica NASE (T) con GUA (T)=Bajo

Del análisis realizado a los proyectos que más inciden cuando la métrica GUA (T) toma un valor BAJO, se concluye que cuando el número de atributos útiles sin errores y útiles con defectos no superan el 20%, y el número de atributos no correctos y no significativos está entre 60 y 100%, se obtiene la mayor cantidad de proyectos con características similares y un valor BAJO para la métrica GUA (T). Esto representa el 63% de los proyectos simulados con un grado BAJO de utilidad de atributos, obteniéndose un valor medio relativo para la métrica GUA (T) del 14%.

d) Regla Experimental

Como se mencionó en la sub-sección *5.b*, al iniciar la etapa de preparación de los datos para un proyecto de explotación de información, se necesita contar con una alta cantidad de atributos útiles en la/s tabla/s disponibles, de manera de favorecer la viabilidad en la realización del proyecto y facilitar además, la selección de aquellos atributos necesarios para generar los modelos.

TABLAXLIV. PROYECTOS CON MAYOR INCIDENCIA SOBRE GUA (T)=BAJO RESPECTO DE LAS MÉTRICAS NASE (T) Y NAUD (T)

		% Útiles sin Errores - NASE (T)				
		0	***10***	***20***	***30***	***Total***
% Útiles con Defectos NAUD (T)	***0***	**880**	**800**	**720**	640	**3040**
	10	**800**	**720**	**640**	0	**2160**
	20	**720**	**640**	**560**	0	**1920**
	30	**640**	**560**	0	0	**1200**
	40	**560**	**480**	0	0	**1040**
	50	480	0	0	0	**480**
	60	400	0	0	0	**400**
	Total	**4480**	**3200**	**1920**	**640**	***10240***

Como regla experimental del comportamiento de la métrica analizada, se concluye que:

- El grado de utilidad de los atributos de la/s tabla/s del proyecto es ALTO, cuando se dispone de un mínimo de 80% de atributos útiles totales (rango de atributos "muchos"), asociado a la métrica NAU (T), que representa una cantidad de atributos útiles sin errores, asociado a la métrica NASE (T), mayor o igual al 70% y una cantidad de atributos útiles con defectos, asociado a la métrica NAUD (T), menor al 30%. A su vez, la cantidad de atributos no útiles, asociado a la métrica NANU (T) y representado por las métricas NANC (T) y NANS (T), no debe superar el 20% del total de atributos (rango de atributos "muy pocos" a "pocos"). En la Fig. 11 se muestra la proporción de atributos necesarios en la/s tabla/s por cada métrica analizada. Con esta regla de comportamiento se obtiene una exactitud de la métrica GUA (T) de un 75% en la categorización de los proyectos con ALTO grado de utilidad de atributos.
- El grado de utilidad de los atributos de la/s tabla/s del proyecto es MEDIO, cuando la cantidad de atributos útiles, asociado a la métrica NAU (T), está entre el 40 y 80% del total (rango de atributos "normal" a "muchos"), que representa una cantidad de atributos útiles sin errores, asociado a la métrica NASE (T), entre el 30 y 70% y una

cantidad de atributos útiles con defectos, asociado a la métrica NAUD (T), menor al 50%. A su vez, la cantidad de atributos no útiles, asociado a la métrica NANU (T) y representado por las métricas NANC (T) y NANS (T), debe ser entre el 20 y 60% del total de atributos (rango de atributos "pocos" a "bastantes"). En la Fig. 11 se muestra la proporción de atributos necesarios en la/s tabla/s por cada métrica analizada. Con esta regla de comportamiento se obtiene una exactitud de la métrica GUA (T) de un 56% en la categorización de los proyectos grado de utilidad de atributos MEDIO.

El grado de utilidad de los atributos de la/s tabla/s del proyecto es BAJO, cuando se dispone de un máximo de 40% de atributos útiles totales (rango de atributos "muy pocos" a "normal"), asociado a la métrica NAU (T), que representa una cantidad de atributos útiles sin errores, asociado a la métrica NASE (T), menor al 30% y una cantidad de atributos útiles con defectos, asociado a la métrica NAUD (T), menor o igual al 20%. A su vez, la cantidad de atributos no útiles, asociado a la métrica NANU (T) y representado por las métricas NANC (T) y NANS (T), debe ser como mínimo el 60% del total de atributos (rango de atributos "bastantes" a "muchos"). En la Fig. 11 se muestra la proporción de atributos necesarios en la/s tabla/s por cada métrica analizada. Con esta regla de comportamiento se obtiene una exactitud de la métrica GUA (T) de un 63% en la categorización de los proyectos con BAJO grado de utilidad de atributos.

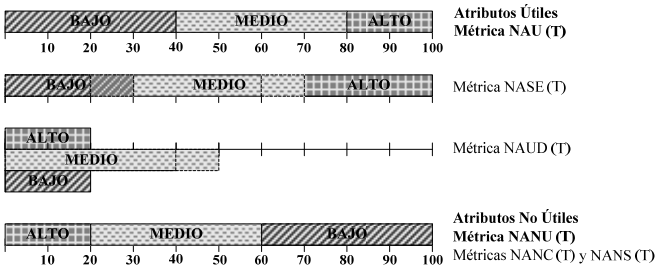

Fig. 11. Proporción de Atributos Útiles y No Útiles según rango en Métrica GUA (T)

Se ha planteado una primera definición de una métrica para medir el grado de utilidad de los atributos de la/s tabla/s en la etapa de preparación de los datos de un proyecto de explotación de información. Para ello, se han propuesto cuatro métricas básicas (variables independientes) que influyen en la utilidad que presenta/n la/s tabla/s para el proyecto. A partir de esta definición de la métrica y de los parámetros y valores establecidos para medir su comportamiento, se concluye que la misma presenta una precisión promedio del 65% en la categorización del grado de utilidad de los atributos de la/s tabla/s de un proyecto.

La métrica y los parámetros iniciales propuestos pueden ser refinados para obtener una mayor precisión de su comportamiento. No obstante, se dispone de una primera métrica con valores de referencia que puede ser utilizada como técnica para medir el grado de utilidad de una o varias tablas para un proyecto de explotación de información.

D. Estudio de las Métricas de Modelos

Para estudiar el comportamiento de las Métricas de Modelos, se decide generar un banco de pruebas simulado con diferentes cantidades de modelos de explotación de información, considerando las restricciones indicadas en [65] para un proyecto de tamaño pequeño, al que se le aplican las métricas propuestas para Modelado (apartado IV, sección *B*, sub-sección *2.a*) cubriendo además, los procesos de explotación de información definidos por [5] para su construcción: Modelo de Descubrimiento de Grupos (apartado IV, sección *B*, sub-sección *2.b*), Modelo de Descubrimiento de Reglas (apartado IV, sección *B*, sub-sección *2.c*) y Modelo de Descubrimiento de Dependencias Significativas (apartado IV, sección *B*, sub-sección *2.d*).

La simulación de las Métricas de Modelos utiliza las siguientes variables independientes (sub-sección *1*) y dependientes (sub-sección *2*).

1. Variables Independientes

Las variables independientes que se van a generar mediante el proceso de simulación, son las correspondientes a las métricas básicas definidas en la sub-secciones de Modelado (apartado IV, sección *B*, sub-sección *2.a*), Modelos de Descubrimiento de Grupos (apartado IV, sección *B*, sub-sección *2.b*), Reglas (apartado IV, sección *B*, sub-sección *2.c*) y Dependencias Significativas (apartado IV, sección *B*, sub-sección *2.d*) y que afectan directamente a las métricas derivadas. Para estas métricas básicas se define un valor específico o un valor aleatorio, considerando las restricciones por el tamaño del proyecto, restringiendo así la cantidad de combinaciones. Las variables independientes a ser utilizadas se muestran en la Tabla XLVI.

2. Variables Dependientes

Para este proceso de simulación las variables dependientes, o sea las que son afectadas por las variables independientes, son los resultados de aplicar las fórmulas de las métricas derivadas de Exactitud del Modelo (sub-sección *3*), Precisión del Modelo (sub-sección *4*), Tasas de Aciertos y Tasa Errores del Modelo (sub-sección *5*), Cobertura de una Regla (sub-sección *6*), Precisión de una Regla (sub-sección *7*), Usabilidad de los Atributos del Modelo (sub-sección *8*),y Grado de Incidencia de un Atributo (sub-sección *9*), para las variables independientes definidas. Las variables dependientes a ser utilizadas se muestran en la Tabla XLVII.

Cabe mencionar, que algunas de las métricas de modelos sólo se calculan en función de la suma de los valores de otras métricas básicas y no se requiere estudiar su comportamiento. Estas métricas denominadas por la norma ISO/IEC 9126 como de agregación se indican en la Tabla XLVIII. La relación entre las variables independientes y dependientes indicando como afectan unas a otras puede verse en la Fig. 12.

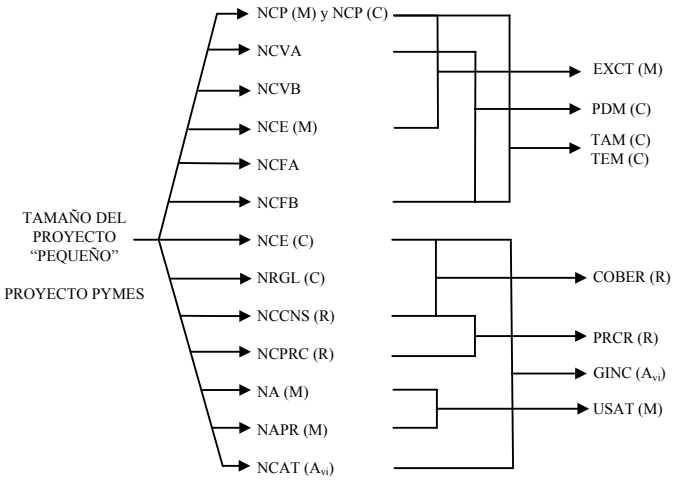

Fig. 12. Relación entre las Variables Independientes y Dependientes para Métricas de Modelos

TABLA XLV. PROYECTOS CON MAYOR INCIDENCIA SOBRE GUA (T)=BAJO RESPECTO DE LAS MÉTRICAS NANC (T) Y NANS (T)

		% No Correctos - NANC (T)											
		0	***10***	***20***	***30***	***40***	***50***	***60***	***70***	***80***	***90***	***100***	***Total***
% No Significativos NANS (T)	***0***	0	0	0	0	0	0	**80**	**160**	**240**	**160**	**80**	**720**
	10	0	0	0	0	0	**80**	**160**	**240**	**160**	**80**	0	**720**
	20	0	0	0	0	**80**	**160**	**240**	**160**	**80**	0	0	**720**
	30	0	0	0	**80**	**160**	**240**	**160**	**80**	0	0	0	**720**
	40	0	0	**80**	**160**	**240**	**160**	**80**	0	0	0	0	**720**
	50	0	**80**	**160**	**240**	**160**	**80**	0	0	0	0	0	**720**
	60	**80**	**160**	**240**	**160**	**80**	0	0	0	0	0	0	**720**
	70	**160**	**240**	**160**	**80**	0	0	0	0	0	0	0	**640**
	80	**240**	**160**	**80**	0	0	0	0	0	0	0	0	**480**
	90	**160**	**80**	0	0	0	0	0	0	0	0	0	**240**
	100	**80**	0	0	0	0	0	0	0	0	0	0	**80**
	Total	**720**	**720**	**720**	**720**	**720**	**720**	**720**	**640**	**480**	**240**	**80**	***6480***

TABLA XLVI. VARIABLES INDEPENDIENTES PARA MÉTRICAS DE MODELOS

Variable Independiente (métrica básica)	**Descripción**
NA (M)	Número de atributos utilizados para construir el modelo, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NAPR (M)	Número de atributos distintos sobre los cuales las reglas de pertenencia a cada clase o grupo del modelo imponen condiciones, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCAT (A_{Vi})	Número de casos cubiertos por una clase o grupo del atributo clase, cuando un atributo significativo A toma el valor V_i, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCCNS (R)	Número de casos que satisfacen la regla de comportamiento o de pertenencia de una clase o grupo, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCE (C)	Número de casos a utilizar por cada clase o grupo del modelo para su entrenamiento, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCE (M)	Número de casos a utilizar para el entrenamiento del modelo, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCFA	Número de casos que fueron incorrectamente clasificados por el modelo como de clase A, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCFB	Número de casos que fueron incorrectamente clasificados por el modelo como de clase A, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCP (C)	Número de casos a utilizar por cada clase o grupo para la prueba del modelo, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCP (M)	Número de casos a utilizar para la prueba del modelo, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCPRC (C)	Número de casos que satisfacen la precondición de la regla, independientemente de la clase o grupo a la que pertenece, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCVA	Número de casos pertenecientes a la clase A que fueron correctamente clasificados por el modelo en esa misma clase, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NCVB	Número de casos pertenecientes a la clase B que fueron correctamente clasificados por el modelo en esa misma clase, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NRGL (C)	Número de reglas descubiertas por del modelo por cada clase o grupo.

TABLA XLVII. VARIABLES DEPENDIENTES PARA MÉTRICAS DE MODELOS

Variable Dependiente (métrica derivada)	Descripción
COBER (C)	Cobertura o soporte de una regla de pertenencia a una clase o grupo, la cual depende del número de casos de entrenamiento de una clase o grupo y del número de casos que satisfacen la regla de comportamiento o de pertenencia a esa clase.
EXCT (M)	Exactitud del modelo para clasificar clases o grupos, la cual depende del número de clasificaciones correctas que tenga un modelo, respecto del número total de casos del mismo.
GINC (A_{Vi})	Grado de incidencia que cada valor V_i de un atributo significativo tiene sobre una clase o grupo del atributo clase, la cual depende del número de casos de entrenamiento que son cubiertos para cada clase o grupo del atributo clase, cuando un atributo significativo toma el valor V_i, respecto del total de casos que pertenecen a esa misma clase.
PDM (C)	Precisión o confianza del modelo para clasificar clases o grupos, la cual depende del número de clasificaciones correctas en una clase o grupo respecto del número de clasificaciones incorrectas de casos en esa misma clase o grupo.
PRCR (C)	Precisión o confianza de una regla de asociación a una clase, la cual depende del número de casos que satisfacen la precondición de una regla (independientemente de la clase), respecto del número de casos que satisfacen la aplicación de esa regla de pertenencia a una clase o grupo.
TAM (C)	Tasa de aciertos del modelo, la cual depende del número de casos de prueba que fueron correctamente clasificados en su clase o grupo, respecto del número de casos de esa clase que fueron incorrectamente clasificados en otras clases o grupos.
TEM (C)	Tasa de errores del modelo, la cual depende del resultado obtenido en la métrica TAM (C), ya que se interpreta como complemento de ésta.
USAT (M)	Usabilidad de los atributos del modelo, la cual depende del número total de atributos distintos utilizados para generar las reglas de pertenencia de clases o grupos del modelo, respecto del número de atributos utilizados para construir dicho modelo.

TABLA XLVIII. VARIABLES DEPENDIENTES PARA MÉTRICAS DE MODELOS NO CONSIDERADAS EN EL ANÁLISIS

Variable Dependiente (métrica agregación)	Descripción
NCG (G)	Número de casos asignados a cada grupo del modelo, el cual se cumple que: $NTC(M) = \sum_{i=1}^{NGR(M)} NCG(G_i)$
NTC (M)	Número total de casos a utilizar para el modelo, el cual se define como: $NTC(M) = NCE(M) + NCP(M)$ A mayor número de casos, mayor es el tiempo requerido para obtener un modelo entrenado, probado y de calidad [47] para el proyecto de explotación de información.

3. *Métrica de Exactitud del Modelo*

Para estudiar la métrica de Exactitud del Modelo – EXCT (M), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de la misma.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en la métrica EXCT (M), considerada como métrica derivada. La lista de variables experimentales se muestra en la Tabla XLIX.

En la sección de diseño experimental (sub-sección *b*) se establece como criterio para la simulación, los valores asociados al número de casos a utilizar en la fase de entrenamiento del modelo del proyecto de explotación de información y no los valores correspondientes a los casos utilizados en la fase prueba del mismo. Por consiguiente, la métrica NCP (M) no se considera para este análisis, sino que se toma como referencia la métrica NCE (M). No obstante, los resultados experimentales y conclusiones obtenidas son de aplicación en ambas fases (entrenamiento y prueba) y con ambas métricas, ya que se corresponden con el número total de casos a utilizar.

TABLA XLIX. VARIABLES EXPERIMENTALES PARA LA MÉTRICA EXCT (M)

Variable Experimental	Descripción
EXCT (M)	Exactitud del modelo para clasificar clases o grupos, la cual se define como: $EXCT(M) = \frac{NCVA + NCVB}{NCE(M)}$ en la fase de entrenamiento del modelo $EXCT(M) = \frac{NCVA + NCVB}{NCP(M)}$ en la fase de prueba del modelo
NCVA	Número de casos pertenecientes a la clase A que fueron correctamente clasificados por el modelo en esa misma clase.
NCVB	Número de casos pertenecientes a la clase B que fueron correctamente clasificados por el modelo en esa misma clase. Para un problema de N clases, representa todos aquellos casos que fueron clasificados correctamente en otras clases diferentes de la clase A.
NCE (M)	Número total de casos a utilizar para el entrenamiento del modelo.
NCP (M)	Número total de casos a utilizar para la prueba del modelo.

b) Diseño Experimental

Para analizar el comportamiento de la métrica enunciada, se utiliza un banco de pruebas simulado de 2.460 modelos de explotación de información, definiendo un rango de valores específico para cada una de las variables experimentales independientes y considerando las restricciones indicadas en [65] según el tamaño del proyecto. De esta manera, se generan diferentes combinaciones sujetas a análisis. Los valores simulados para las variables experimentales independientes se muestran en la Tabla L.

El resultado y comportamiento de la métrica derivada EXCT (M) está asociado al número de casos correctamente clasificados en su correspondiente clase o grupo dentro del modelo de explotación de información construido, cuyas métricas básicas NCVA y NCVB, se indican en la Tabla L. Por otra parte, se tiene como restricción que la cantidad de casos clasificados correcta e incorrectamente deben cubrir la totalidad de los casos utilizados en el modelo construido:

$$NCE(M) = NCVA + NCVB + Nro\_Clasificaciones\_Incorrectas$$

TABLA L. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LA MÉTRICA EXCT (M)

Variable Experimental	Descripción
NCVA	Número de casos pertenecientes a la clase A que fueron correctamente clasificados por el modelo en esa misma clase, con un rango de valores específico 0, 10, 20, 30, 40, 50, 60, 70, 80 y 90. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en el modelo.
NCVB	Número de casos pertenecientes a la clase B que fueron correctamente clasificados por el modelo en esa misma clase, con un rango de valores específico 0, 10, 20, 30, 40, 50, 60, 70, 80 y 90. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en el modelo.
NCE (M)	Número de casos a utilizar para el entrenamiento del modelo, con un rango de valores específico 1.000.000, 2.000.000, 3.000.000, 4.000.000, 5.000.000, 6.000.000, 7.000.000, 8.000.000, 9.000.000 y 10.000.000, según lo predefinido para un proyecto de tamaño pequeño.

Al tratarse de métricas básicas sometidas a un proceso de simulación, no se les puede asignar un número a priori de casos a utilizar. Por tal motivo, se decide establecer un rango de valores porcentuales con intervalos de a 10% (entre 0 y 90%), respecto del número total de casos del modelo, con el fin de analizar el comportamiento general de la métrica EXCT (M), según estos rangos, tal como se indica en la Tabla LI. Por otra parte, y como se indicó anteriormente, existe la restricción entre estas métricas respecto del total de casos del modelo. Por consiguiente, se necesitan sólo aquellas configuraciones de las métricas NCVA, NCVB y el número de clasificaciones incorrectas, cuya suma represente el 100% de los casos. De este análisis se obtienen 246 configuraciones posibles aplicables a un modelo.

La exactitud de un modelo de clasificación de casos en clases o grupos es muy importante al momento de determinar la calidad de un clasificador, y se suele estimar tras ejecutar el modelo construido sobre diferentes conjuntos de datos [56]. Sin embargo, no existe un criterio universal que defina qué indicador se debe adoptar para considerar un modelo con ALTA o BAJA exactitud de clasificación en un proyecto de explotación de información. Por consiguiente, y a partir de la propia experiencia, en la Tabla LII se indican los valores y rangos contemplados para un proyecto de explotación de información para PyMEs.

TABLA LI. RANGOS Y VALORES RELACIONADOS A LAS MÉTRICAS NCVA Y NCVB PARA CLASIFICACIONES CORRECTAS

Métrica	Descripción	Valores de simulación
NCVA NCVB	Ningún caso clasificado correctamente	0
	De 1 a 10% de casos clasificados correctamente	10
	De 11 a 20% de casos clasificados correctamente	20
	De 21 a 30% de casos clasificados correctamente	30
	De 31 a 40% de casos clasificados correctamente	40
	De 41 a 50% de casos clasificados correctamente	50
	De 51 a 60% de casos clasificados correctamente	60
	De 61 a 70% de casos clasificados correctamente	70
	De 71 a 80% de casos clasificados correctamente	80
	De 81 a 90% de casos clasificados correctamente	90

A partir de la asignación de los valores a cada variable independiente, se generan los datos de los modelos de explotación de información. De esta manera se busca obtener los resultados y el comportamiento de la métrica EXCT (M) (variable experimental dependiente) en función de la variación de estas variables.

TABLA LII. RANGOS Y VALORES RELACIONADOS A LA MÉTRICA EXCT (M)

Rango	Descripción	Valores de simulación
Bajo	De 0 a 69% de casos clasificados correctamente	Bajo
Alto	Más de 70% de casos clasificados correctamente	Alto

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de la métrica EXCT (M), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que la métrica derivada EXCT (M) toma un valor ALTO, considerándose como parámetro para este análisis un valor relativo superior al 70%, cuando el porcentaje de casos clasificados correctamente para las clases A y B, asociado a las métricas NCVA y NCVB, está entre 0 y 90% del total de casos utilizados en el modelo. En la Tabla LIII, se muestra la distribución del número de modelos en función del porcentaje de casos clasificados correctamente. En esta tabla, se observa que para que el grado de exactitud de un modelo sea ALTO se necesita que el mismo sea capaz de clasificar correctamente en cada clase o grupo, como mínimo el 70% de los casos totales.

Con esta información, se decide estudiar el comportamiento de aquellos modelos que mejor caracterizan el grado de la métrica EXCT (M) con valor ALTO.

Del análisis experimental, se observa que la métrica EXCT (M) instancia una mayor cantidad de modelos con valor ALTO, cuando el porcentaje de casos clasificados correctamente para la clase A, asociado a la métrica NCVA, está entre 10 y 70% del total de casos del modelo.

A su vez, el porcentaje de casos clasificados correctamente en otras clases, asociado a la métrica NCVB, también se ubica dentro del mismo valor de casos totales. La variación de los rangos de esta métrica respecto al total de modelos simulados se muestra en la Fig. 13. Esta proporción de clasificaciones correctas en cada clase o grupo representa el 82% de los modelos simulados con ALTO grado de exactitud de clasificación, obteniéndose un valor medio relativo para la métrica EXCT (M) del 80%.

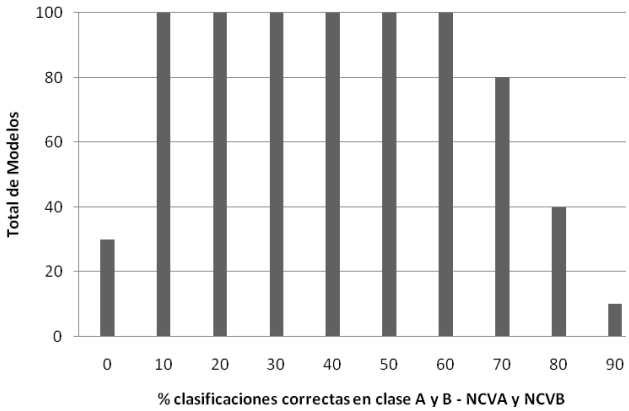

Fig. 13. Variación de rangos de la métrica NCVA y NCVB con EXCT (M) = Alto

TABLA LIII. DISTRIBUCIÓN DE MODELOS CON VALOR EXCT (M)=ALTO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)										
		0	***10***	***20***	***30***	***40***	***50***	***60***	***70***	***80***	***90***	***Total***
% Clasificaciones como clase B correctas - (NCVB)	***0***	0	0	0	0	0	0	0	20	10	0	**30**
	10	0	0	0	0	0	0	40	30	20	10	**100**
	20	0	0	0	0	0	40	30	20	10	0	**100**
	30	0	0	0	0	40	30	20	10	0	0	**100**
	40	0	0	0	40	30	20	10	0	0	0	**100**
	50	0	0	40	30	20	10	0	0	0	0	**100**
	60	0	40	30	20	10	0	0	0	0	0	**100**
	70	20	30	20	10	0	0	0	0	0	0	**80**
	80	10	20	10	0	0	0	0	0	0	0	**40**
	90	0	10	0	0	0	0	0	0	0	0	**10**
	Total	**30**	**100**	**100**	**100**	**100**	**100**	**100**	**80**	**40**	**10**	***760***

En la Tabla LIV, se indican aquellos modelos con mayor incidencia en la métrica EXCT (M) en función del porcentaje de casos clasificados correctamente en cada clase o grupo.

En esta tabla se observa además, que si el porcentaje de clasificaciones correctas en alguna de las clases es superior al 70%, decrece considerablemente (hasta inclusive 0%) el número de clasificaciones correctas de otras clases. Esto implica, que si bien el grado de exactitud del modelo es ALTO, no es un buen modelo para clasificar correctamente los casos en todas las clases o grupos del modelo.

En el trabajo de [38] se estudia además, cómo a medida que aumenta el número de clases o grupos utilizados en la construcción del modelo, la efectividad del mismo decrece.

Por otra parte, en [22] se menciona que si el número total de casos a utilizar en la construcción de un modelo presenta un número de casos por clase aproximadamente igual, el modelo puede producir clasificaciones más acertadas. Sin embargo, en muchos proyectos reales, no se cumple la hipótesis de contar con un conjunto de casos uniformemente distribuidos entre las clases, ocasionando que las clases queden desbalanceadas o sesgadas. El desbalance en un conjunto de casos se refiere a la situación en la que una clase (mayoritaria) presenta una cantidad notablemente mayor de casos en comparación con otra clase (minoritaria).

d) Regla Experimental

Como regla experimental del comportamiento de la métrica analizada, se concluye que cuando el número de casos totales clasificados correctamente por el modelo de explotación de información es superior al 70% de los casos totales utilizados, el número de casos clasificados correctamente para cada una de las clases o grupos está entre 10 y 70% de los casos totales (equivale de 1 a 70% de casos), y existe una distribución balanceada de casos por clase, se obtiene la mayor cantidad de modelos con características similares y un valor ALTO para la métrica EXCT (M), considerándose un buen modelo de explotación de información para clasificar nuevos casos del mismo dominio.

Caso contrario, si el valor de la métrica EXCT (M) es inferior al 70% de clasificaciones correctas, no se lo puede considerar un buen modelo para clasificar nuevos casos, ya que produce un aumento en el número de clasificaciones incorrectas.

4. Métrica de Precisión del Modelo

Para estudiar la métrica de Precisión del Modelo para clasificar una clase – PDM (C), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de la misma.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en la métrica PDM (C), considerada como métrica derivada. La lista de variables experimentales se muestra en la Tabla LV.

TABLA LV. VARIABLES EXPERIMENTALES PARA LA MÉTRICA PDM (M)

Variable Experimental	Descripción
NCVA	Número de casos pertenecientes a la clase A que fueron correctamente clasificados por el modelo en esa misma clase.
NCVB	Número de casos pertenecientes a la clase B que fueron correctamente clasificados por el modelo en esa misma clase. Para un problema de N clases, representa todos aquellos casos que fueron clasificados correctamente en otras clases diferentes de la clase A.
NCFA	Número de casos que fueron incorrectamente clasificados por el modelo como clase A, pero que pertenecen a otra clase.
NCFB	Número de casos que pertenecen a la clase A pero que fueron incorrectamente clasificados por el modelo en otra clase. Para un problema de N clases, representa todos aquellos casos que fueron clasificados incorrectamente en otras clases diferentes de la clase A.
NCE (M)	Número de casos a utilizar para el entrenamiento del modelo.
PDM (C)	Precisión o confianza del modelo para clasificar una clase, la cual se define como: $\mathrm{PDM}(C_i) = \frac{\mathrm{NCVC_i}}{\mathrm{NCVC_i} + \mathrm{NCFC_i}}$ donde C_i es la clase correspondiente que se está analizando.

En la sección de diseño experimental (sub-sección b) se establece como criterio para la simulación, los valores asociados al número de casos a utilizar en la fase de entrenamiento del modelo del proyecto de explotación de información, ya que la precisión de un modelo es un indicador de la efectividad del mismo para clasificar nuevos casos a una clase o grupo en particular, luego de construido y entrenado el modelo.

TABLA LIV. MODELOS CON MAYOR INCIDENCIA SOBRE EXCT (M) =ALTO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)										
		0	***10***	***20***	***30***	***40***	***50***	***60***	***70***	***80***	***90***	***Total***
% Clasificaciones como clase B correctas - (NCVB)	***0***	0	0	0	0	0	0	0	20	10	0	**30**
	10	0	**0**	**0**	**0**	**0**	**0**	**40**	**30**	20	10	**100**
	20	0	**0**	**0**	**0**	**0**	**40**	**30**	**20**	10	0	**100**
	30	0	**0**	**0**	**0**	**40**	**30**	**20**	**10**	0	0	**100**
	40	0	**0**	**0**	**40**	**30**	**20**	**10**	**0**	0	0	**100**
	50	0	**0**	**40**	**30**	**20**	**10**	**0**	**0**	0	0	**100**
	60	0	**40**	**30**	**20**	**10**	**0**	**0**	**0**	0	0	**100**
	70	20	**30**	**20**	**10**	**0**	**0**	**0**	**0**	0	0	**80**
	80	10	20	10	0	0	0	0	0	0	0	**40**
	90	0	10	0	0	0	0	0	0	0	0	**10**
	Total	**30**	**100**	**100**	**100**	**100**	**100**	**100**	**80**	**40**	**10**	***760***

b) Diseño Experimental

Para analizar el comportamiento de la métrica enunciada, se utiliza un banco de pruebas simulado de 760 modelos de explotación de información, definiendo un rango de valores específico para cada una de las variables experimentales independientes y considerando las restricciones indicadas en [65] según el tamaño del proyecto. Además, se considera que la exactitud de estos modelos es alta. De esta manera, se generan diferentes combinaciones sujetas a análisis. Los valores simulados para las variables experimentales independientes se muestran en la Tabla LVI.

El resultado y comportamiento de la métrica derivada PDM (C) está asociado al número de casos reales que pertenecen a una determinada clase o grupo, respecto del total de casos que son clasificados por el modelo de explotación de información para esa misma clase, cuyas métricas básicas NCVA, NCVB, NCFA y NCFB, se indican en la Tabla LVI. A su vez, las primeras dos métricas representan la cantidad total de casos clasificados correctamente por el modelo, mientras que las dos últimas métricas representan la cantidad total de casos clasificados incorrectamente

Por otra parte, se tiene como restricción que la cantidad de casos clasificados correcta e incorrectamente debe cubrir la totalidad de los casos utilizados para el entrenamiento del modelo construido.

$$NCE(M) = NCE(C_A) + NCE(C_B)$$

donde:

- $NCE(C_A) = NCVA + NCFB$ es el número de casos de entrenamiento para la clase A.
- $NCE(C_B) = NCVB + NCFA$ es el número de casos de entrenamiento para la clase B.

Al tratarse de métricas básicas sometidas a un proceso de simulación, no se les puede asignar un número a priori de casos a utilizar. Por tal motivo, se decide establecer un rango de valores porcentuales con intervalos de a 10%, definido entre 0 y 90% para las clasificaciones correctas y entre 0 y 30% para las clasificaciones incorrectas, respecto del número total de casos del modelo, tal como se indican en las Tablas LVII y LVIII. Estos valores, a su vez, comprenden a los modelos de explotación de información cuya métrica de exactitud es superior al 70% (sub-sección 3). A partir de los rangos de valores definidos, se analiza el comportamiento general de la métrica PDM (C).

TABLA LVI. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LA MÉTRICA PDM (M)

Variable Experimental	Descripción
NCVA	Número de casos pertenecientes a la clase A que fueron correctamente clasificados por el modelo en esa misma clase, con un rango de valores específico 0, 10, 20, 30, 40, 50, 60, 70, 80 y 90. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en el modelo.
NCVB	Número de casos pertenecientes a la clase B que fueron correctamente clasificados por el modelo en esa misma clase, con un rango de valores específico 0, 10, 20, 30, 40, 50, 60, 70, 80 y 90. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en el modelo.
NCFA	Número de casos que fueron incorrectamente clasificados por el modelo como clase A, pero que pertenecen a otra clase, con un rango de valores específico 0, 10, 20 y 30. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en el modelo.
NCFB	Número de casos que pertenecen a la clase A pero que fueron incorrectamente clasificados por el modelo en otra clase, con un rango de valores específico 0, 10, 20 y 30. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en el modelo.
NCE (M)	Número de casos a utilizar para el entrenamiento del modelo, con un rango de valores específico 1.000.000, 2.000.000, 3.000.000, 4.000.000, 5.000.000, 6.000.000, 7.000.000, 8.000.000, 9.000.000 y 10.000.000, según lo predefinido para un proyecto de tamaño pequeño.

Como se indicó anteriormente, existe la restricción entre estas métricas respecto del total de casos del modelo. Por consiguiente, se necesitan sólo aquellas configuraciones de las métricas NCVA, NCVB, NCFA y NCFB, cuya suma represente el 100% de los casos. De este análisis se obtienen 76 configuraciones posibles aplicables a un modelo.

En virtud que no existe un criterio universal que defina qué indicador se debe adoptar para considerar un modelo con ALTA o BAJA precisión para clasificar una clase en un proyecto de explotación de información, en la Tabla LVIX se indica los valores y rangos contemplados para un proyecto de explotación de información para PyMEs, considerados a partir de la propia experiencia.

TABLA LVII. RANGOS Y VALORES RELACIONADOS A LAS MÉTRICAS NCVA Y NCVB PARA CLASIFICACIONES CORRECTAS

Métrica	Descripción	Valores de simulación
NCVA NCVB	Ningún caso clasificado correctamente	0
	De 1 a 10% de casos clasificados correctamente	10
	De 11 a 20% de casos clasificados correctamente	20
	De 21 a 30% de casos clasificados correctamente	30
	De 31 a 40% de casos clasificados correctamente	40
	De 41 a 50% de casos clasificados correctamente	50
	De 51 a 60% de casos clasificados correctamente	60
	De 61 a 70% de casos clasificados correctamente	70
	De 71 a 80% de casos clasificados correctamente	80
	De 81 a 90% de casos clasificados correctamente	90

A partir de la asignación de los valores a cada variable independiente, se generan los datos de los modelos de explotación de información. De esta manera se busca obtener los resultados y el comportamiento de la métrica PDM (C) (variable experimental dependiente) en función de la variación de estas variables.

TABLA LVIII. RANGOS Y VALORES RELACIONADOS A LAS MÉTRICAS NCFA Y NCFB PARA CLASIFICACIONES INCORRECTAS

Métrica	Descripción	Valores de simulación
NCFA NCFB	Ningún caso clasificado incorrectamente	0
	De 1 a 10% de casos clasificados incorrectamente	10
	De 11 a 20% de casos clasificados incorrectamente	20
	De 21 a 30% de casos clasificados incorrectamente	30

TABLA LVIX. RANGOS Y VALORES RELACIONADOS A LA MÉTRICA PDM (M)

Rango	Descripción
Bajo	Menos de 50% de efectividad para clasificar una clase
Medio	Entre 50 y 70% de efectividad para clasificar una clase
Alto	Más de 70% de efectividad para clasificar una clase

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de la métrica PDM (C), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que la métrica derivada PDM (C) toma un valor ALTO, considerándose como parámetro para este análisis un valor relativo superior al 70%, cuando el porcentaje de casos clasificados correctamente para las clases A y B, asociado a las métricas NCVA y NCVB, está entre 10 y 90% del total de casos utilizados en el modelo. En la Tabla LX, se muestra la distribución del número de modelos en función del porcentaje de casos clasificados correctamente. En esta tabla, se observa que para que el grado de precisión de un modelo sea ALTO se necesita que el mismo tenga una exactitud mínima del 80% de casos clasificados correctamente en su clase o grupo.

Por otra parte, el porcentaje de clasificaciones incorrectas en cada una de las clases, asociadas a las métricas NCFA y NCFB, está entre 0 y 20% del total de casos del modelo. En la Tabla LXI, se muestra la distribución del número de modelos en función del porcentaje de casos clasificados incorrectamente en cada clase. En la misma, se observa que para que el grado de precisión de un modelo sea ALTO se necesita un máximo de 20% de casos clasificados incorrectamente en otras clases.

LXI. DISTRIBUCIÓN DE MODELOS CON VALOR PDM (C)=ALTO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase A incorrectas - (NCFA)			
		0	***10***	***20***	***Total***
% Clasificaciones como clase B incorrectas - (NCFB)	***0***	90	60	30	**180**
	10	60	30	0	**90**
	20	30	0	0	**30**
	Total	**180**	**90**	**30**	***300***

Con esta información, se decide estudiar el comportamiento de aquellos modelos que mejor caracterizan el grado de la métrica PDM (C) con valor ALTO.

Del análisis experimental, se observa que la métrica PDM (C) instancia una mayor cantidad de modelos con valor ALTO, cuando el porcentaje de casos clasificados correctamente para la clase A, asociado a la métrica NCVA, está entre 20 y 70% del total de casos del modelo. A su vez, el porcentaje de casos clasificados correctamente en otras clases, asociado a la métrica NCVB, también se ubica dentro del mismo valor de casos totales. Esta proporción de clasificaciones correctas en cada clase o grupo representa el 73% de los modelos simulados con ALTO grado de precisión de clasificación, obteniéndose un valor medio relativo para la métrica PDM (C) del 91%.

En la Tabla LXII, se indican aquellos modelos con mayor incidencia en la métrica PDM (C) en función del porcentaje de casos clasificados correctamente en cada clase o grupo.

Por otra parte, para los modelos que más inciden sobre la métrica PDM (C), se observa que en mayor proporción el porcentaje de casos clasificados incorrectamente, asociados a las métricas NCFA y NCFB, es menor al 20% del total de casos del modelo, tal como se muestra en la Tabla LXIII. Esta proporción de clasificaciones incorrectas representa el 82% de los modelos de explotación de información con mayor incidencia sobre la métrica en estudio, mientras que el 18% restante presenta un comportamiento extrapolado respecto a los casos clasificados incorrectamente para cada clase.

Del análisis realizado a los modelos que más inciden cuando la métrica PDM (C) toma un valor ALTO, se concluye que cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos está entre 20 y 70% del total de casos, el número de casos clasificados incorrectamente en esa misma clase está entre 0 y 10% y la exactitud del modelo es superior al 80%, se obtiene la mayor cantidad de modelos de explotación de información con características similares y un valor ALTO para la métrica PDM (C). Esto representa el 60% de los modelos simulados con ALTO grado de precisión para clasificar una clase, obteniéndose un valor medio relativo para la métrica PDM (C) del 92%. Sin embargo, cuando el número de casos clasificados incorrectamente impacta sólo en una de las clases (una de las métricas toma valor 20% y la otra 0%), manteniendo el número de casos clasificados correctamente, el valor medio relativo para la métrica PDM (C) desciende al 73%, lo que representa el 13% de los modelos simulados.

TABLA LX. DISTRIBUCIÓN DE MODELOS CON VALOR PDM (C)=ALTO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)									
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	*10*	0	0	0	0	0	0	10	10	10	**30**
	20	0	0	0	0	0	10	10	10	0	**30**
	30	0	0	0	0	20	20	10	0	0	**50**
	40	0	0	0	10	20	10	0	0	0	**40**
	50	0	0	20	20	10	0	0	0	0	**50**
	60	0	10	20	10	0	0	0	0	0	**40**
	70	10	10	10	0	0	0	0	0	0	**30**
	80	10	10	0	0	0	0	0	0	0	**20**
	90	10	0	0	0	0	0	0	0	0	**10**
	Total	**30**	**30**	**50**	**40**	**50**	**40**	**30**	**20**	**10**	***300***

TABLA LXII. MODELOS CON MAYOR INCIDENCIA SOBRE PDM (C)=ALTO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)									
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	*10*	0	**0**	**0**	**0**	**0**	**0**	10	10	10	**30**
	20	0	**0**	**0**	**0**	**0**	**10**	**10**	10	0	**30**
	30	0	**0**	**0**	**0**	**20**	**20**	**10**	0	0	**50**
	40	0	**0**	**0**	**10**	**20**	**10**	**0**	0	0	**40**
	50	0	**0**	**20**	**20**	**10**	**0**	**0**	0	0	**50**
	60	0	**10**	**20**	**10**	**0**	**0**	**0**	0	0	**40**
	70	10	**10**	**10**	**0**	**0**	**0**	**0**	0	0	**30**
	80	10	10	0	0	0	0	0	0	0	**20**
	90	10	0	0	0	0	0	0	0	0	**10**
	Total	**30**	**30**	**50**	**40**	**50**	**40**	**30**	**20**	**10**	***300***

TABLA LXIII. MODELOS CON MAYOR INCIDENCIA SOBRE PDM (C)=ALTO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase A incorrectas - (NCFA)			
		0	*10*	*20*	*Total*
% Clasificaciones como clase B incorrectas - (NCFB)	*0*	**50**	**50**	20	**120**
	10	**50**	**30**	0	**80**
	20	20	0	0	**20**
	Total	**120**	**80**	**30**	***220***

En un segundo análisis experimental, se observa que la métrica derivada PDM (C) toma un valor MEDIO, considerándose como parámetro para este análisis un valor relativo entre 50 y 70%, cuando el porcentaje de casos clasificados correctamente para las clases A y B, asociado a las métricas NCVA y NCVB, está entre 10 y 80% del total de casos utilizados en el modelo. Sin embargo, se presentan dos situaciones que afectan la precisión de cada clase en forma individual y que dependen del número de casos clasificados correctamente en cada una de las clases o grupos.

En la primera de ellas, se observa que la precisión del modelo para clasificar casos en la clase A toma un valor medio mientras la precisión de las otras clases del modelo, correspondiente a la clase B, toma un valor alto. La segunda situación se da al contrario, observándose que la precisión del modelo para clasificar casos en la clase A toma un valor alto, mientras que la precisión de alguna de las otras clases toma un valor medio.

A los fines del análisis del comportamiento de la métrica y de las implicancias para el proyecto de explotación de información, se asume que cuando la precisión en alguna de las clases del modelo construido toma un valor MEDIO, la precisión global del modelo es de valor MEDIO. En las Tablas LXIV y LXV, se muestra la distribución del número de modelos en función del porcentaje de casos clasificados correctamente y de la precisión obtenida para cada clase del modelo. En las mismas, se observa que para que el grado de precisión de un modelo sea MEDIO se necesita que el mismo tenga una exactitud entre 70 y 90% de casos clasificados correctamente en su clase o grupo.

Por otra parte, cuando la precisión del modelo para clasificar casos en la clase A toma un valor MEDIO, el porcentaje de clasificaciones incorrectas para la clase A, asociada a la métrica NCFA, está entre 10 y 30% del total de casos, mientras que el porcentaje total de clasificaciones incorrectas en otras clases del modelo, asociada a la métrica NCFB, es menor al 20% del total de casos. Por el contrario, cuando la precisión para clasificar casos en alguna de las otras clases distintas de la clase A toma un valor MEDIO, el porcentaje de clasificaciones incorrectas para esa clase,

asociada a la métrica NCFB, está entre 10 y 30% del total de los casos, mientras que el porcentaje total de clasificaciones incorrectas para la clase A, asociada a la métrica NCFA, es menor al 20% del total de casos.

En las Tablas LXVI y LXVII, se muestra la distribución del número de modelos en función del porcentaje de casos clasificados incorrectamente y de la precisión obtenida para cada clase del modelo. En las mismas, se observa que para que el grado de precisión de un modelo sea MEDIO se necesita entre 10 y 30% de casos clasificados incorrectamente en el modelo. Con esta información, se decide estudiar el comportamiento de aquellos modelos que mejor caracterizan el grado de la métrica PDM (C) con valor MEDIO.

TABLA LXIV. DISTRIBUCIÓN DE MODELOS CON VALOR PDM (CLASE A)= MEDIO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)								
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	***10***	0	0	0	0	0	10	0	0	**10**
	20	0	0	0	0	10	0	0	0	**10**
	30	0	0	0	20	0	0	0	0	**20**
	40	0	0	20	10	0	0	0	0	**30**
	50	0	20	10	0	0	0	0	0	**30**
	60	10	20	0	0	0	0	0	0	**30**
	70	10	10	0	0	0	0	0	0	**20**
	80	10	0	0	0	0	0	0	0	**10**
	Total	**30**	**50**	**30**	**30**	**10**	**10**	**0**	**0**	***160***

TABLA LXV. DISTRIBUCIÓN DE MODELOS CON VALOR PDM (CLASE B)= MEDIO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)								
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	***10***	0	0	0	0	0	10	10	10	**30**
	20	0	0	0	0	20	20	10	0	**50**
	30	0	0	0	20	10	0	0	0	**30**
	40	0	0	20	10	0	0	0	0	**30**
	50	0	10	0	0	0	0	0	0	**10**
	60	10	0	0	0	0	0	0	0	**10**
	70	0	0	0	0	0	0	0	0	**0**
	80	0	0	0	0	0	0	0	0	**0**
	Total	**10**	**10**	**20**	**30**	**30**	**30**	**20**	**10**	***160***

TABLA LXVI. DISTRIBUCIÓN DE MODELOS CON VALOR PDM (CLASE A)= MEDIO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase A incorrectas - (NCFA)			
		0	*10*	*20*	*Total*
% Clasificaciones como clase B incorrectas - (NCFB)	***0***	20	30	40	**90**
	10	20	30	0	**50**
	20	20	0	0	**20**
	Total	**60**	**60**	**40**	***160***

TABLA LXVII. DISTRIBUCIÓN DE MODELOS CON VALOR PDM (CLASE B)= MEDIO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase A incorrectas - (NCFA)			
		0	*10*	*20*	*Total*
% Clasificaciones como clase B incorrectas - (NCFB)	***0***	20	20	20	**60**
	10	30	30	0	**60**
	20	40	0	0	**40**
	Total	**90**	**50**	**20**	***160***

Del análisis experimental, se observa que la métrica PDM (C) instancia una mayor cantidad de modelos con valor MEDIO en la precisión para la clase A, cuando el porcentaje de clasificaciones correctas para esta clase, asociada a la métrica NCVA, está entre 10 y 40% del total de casos, mientras que el porcentaje total de clasificaciones correctas en otras clases del modelo, asociada a la métrica NCVB, está entre 30 y 70% del total de casos.

Por el contrario, cuando la precisión para clasificar casos en alguna de las otras clases distintas de la clase A toma un valor MEDIO, el porcentaje de clasificaciones correctas para esa clase, asociada a la métrica NCVB, está entre 10 y 40% del total de los casos, mientras que el porcentaje total de clasificaciones correctas para la clase A, asociada a la métrica NCVA, está entre 30 y 70% del total de casos. En ambas situaciones, la proporción de clasificaciones correctas en cada clase o grupo representa el 81% de los modelos simulados con grado MEDIO de precisión de clasificación, obteniéndose un valor medio relativo para la métrica PDM (C) del 58%.

En las Tablas LXVIII y LXIX, se indican aquellos modelos con mayor incidencia en la métrica PDM (C) en función del porcentaje de casos clasificados correctamente y de la precisión obtenida para cada clase o grupo del modelo.

Por otra parte, para los modelos que más inciden sobre la métrica PDM (C) se observa que, cuando la precisión del modelo para clasificar casos en la clase A toma un valor MEDIO, el porcentaje de clasificaciones incorrectas para la clase A, asociada a la métrica NCFA, está entre 10 y 30% del total de casos, mientras que el porcentaje total de clasificaciones incorrectas en otras clases del modelo, asociada a la métrica NCFB, es menor al 20% del total de casos.

TABLA LXVIII. MODELOS CON MAYOR INCIDENCIA SOBRE PDM (CLASE A)= MEDIO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)								
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	***10***	0	0	0	0	0	10	0	0	**10**
	20	0	0	0	0	10	0	0	0	**10**
	30	**0**	**0**	**0**	**20**	0	0	0	0	**20**
	40	**0**	**0**	**20**	**10**	0	0	0	0	**30**
	50	**0**	**20**	**10**	**0**	0	0	0	0	**30**
	60	**10**	**20**	**0**	**0**	0	0	0	0	**30**
	70	**10**	**10**	**0**	**0**	0	0	0	0	**20**
	80	10	0	0	0	0	0	0	0	**10**
	Total	**30**	**50**	**30**	**30**	**10**	**10**	**0**	**0**	***160***

TABLA LXIX. MODELOS CON MAYOR INCIDENCIA SOBRE PDM (CLASE B)= MEDIO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)								
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	***10***	0	0	**0**	**0**	**0**	**10**	**10**	10	**30**
	20	0	0	**0**	**0**	**20**	**20**	**10**	0	**50**
	30	0	0	**0**	**20**	**10**	**0**	**0**	0	**30**
	40	0	0	**20**	**10**	**0**	**0**	**0**	0	**30**
	50	0	10	0	0	0	0	0	0	**10**
	60	10	0	0	0	0	0	0	0	**10**
	70	0	0	0	0	0	0	0	0	**0**
	80	0	0	0	0	0	0	0	0	**0**
	Total	**10**	**10**	**20**	**30**	**30**	**30**	**20**	**10**	***160***

Por el contrario, cuando la precisión para clasificar casos en alguna de las otras clases distintas de la clase A toma un valor MEDIO, el porcentaje de clasificaciones incorrectas para esa clase, asociada a la métrica NCFB, está entre 10 y 30% del total de los casos, mientras que el porcentaje total de clasificaciones incorrectas para la clase A, asociada a la métrica NCFA, es menor al 20% del total de casos, tal como se muestra en las Tablas LXX y LXXI.

Esta proporción de clasificaciones incorrectas representa el 70% de los modelos de explotación de información con mayor incidencia sobre la métrica en estudio, mientras que el 30% restante presenta un comportamiento extrapolado respecto a los casos clasificados incorrectamente para cada clase.

TABLA LXX. MODELOS CON MAYOR INCIDENCIA SOBRE PDM (CLASE A)= MEDIO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase A incorrectas - (NCFA)			
		0	*10*	*20*	*Total*
% Clasificaciones como clase B incorrectas - (NCFB)	***0***	**10**	**30**	20	**60**
	10	**20**	**30**	0	**50**
	20	20	0	0	**20**
	Total	**50**	**60**	**20**	***130***

TABLA LXXI. MODELOS CON MAYOR INCIDENCIA SOBRE PDM (CLASE B)= MEDIO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase A incorrectas - (NCFA)			
		0	*10*	*20*	*Total*
% Clasificaciones como clase B incorrectas - (NCFB)	***0***	**10**	**20**	20	**50**
	10	**30**	**30**	0	**60**
	20	20	0	0	**20**
	Total	**60**	**50**	**20**	***130***

Del análisis realizado a los modelos que más inciden cuando la métrica PDM (C) toma un valor MEDIO, se concluye que cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos está entre 10 y 40% del total de casos, el número de casos clasificados incorrectamente en esa misma clase está entre 10 y 20% aproximadamente y la exactitud del modelo está entre 70 y 90%, se obtiene la mayor cantidad de modelos de explotación de información con características similares y un valor MEDIO para la métrica PDM (C). Esto representa el 57% de los modelos simulados con grado MEDIO de precisión para clasificar una clase, obteniéndose un valor medio relativo para la métrica PDM (C) del 60%. Sin embargo, cuando el número de casos clasificados incorrectamente impacta sólo en una de las clases (una de las métricas toma valor 30% y la otra 0%), manteniendo el número de casos clasificados correctamente, el valor medio relativo para la métrica PDM (C) desciende al 54%, lo que representa el 25% de los modelos simulados.

En el último análisis experimental realizado sobre la métrica PDM (C), se observa que cuando la misma toma un valor BAJO, se presentan las mismas dos situaciones que en el análisis anterior, y que afectan la precisión de cada clase en forma individual dependiendo del número de casos clasificados correctamente en cada una de las clases o grupos. En la primera situación, se observa que la precisión del modelo para clasificar casos en la clase A toma un valor BAJO (la precisión de las otras clases del modelo toma un valor alto) cuando el porcentaje de casos clasificados correctamente para la clase A, asociado a la métrica NCVA, está entre 0 y 20% del total de casos utilizados en el modelo. A su vez, el porcentaje de casos clasificados correctamente en otras clases, correspondiente a la clase B y asociado a la métrica NCVB, está entre 50 y 80%.

Por el contrario, cuando la precisión para clasificar casos en alguna de las otras clases distintas de la clase A toma un valor BAJO, la precisión para clasificar casos en la clase A es alto. En este caso, el porcentaje de clasificaciones correctas para otras clases, asociada a la métrica NCVB, está entre 0 y 20% del total de los casos utilizados en el modelo, mientras que el porcentaje total de clasificaciones correctas para la clase A, asociada a la métrica NCVA, está entre 50 y 80%.

En las Tablas LXXII y LXXIII, se muestra la distribución del número de modelos en función del porcentaje de casos clasificados correctamente y de la precisión obtenida para cada clase del modelo. En las mismas, se observa que para que el grado de precisión de un modelo sea BAJO se necesita que el mismo tenga una exactitud entre 70 y 80% de casos clasificados correctamente en su clase o grupo. A los fines del análisis del comportamiento de la métrica y de las implicancias para el proyecto de explotación de información, se asume que cuando la precisión en alguna de las clases del modelo construido toma un valor BAJO, la precisión global del modelo es de valor BAJO y no se lo debería considerar un buen modelo ya que el mismo, mayormente, confunde los casos al momento de clasificarlos en su clase correspondiente, resultando un modelo poco efectivo para el proyecto de explotación de información.

Por otra parte, cuando la precisión del modelo para clasificar casos en la clase A toma un valor BAJO, el porcentaje de clasificaciones incorrectas para la clase A, asociada a la métrica NCFA, está entre 10 y 30% del total de casos, mientras que el porcentaje total de clasificaciones incorrectas en otras clases del modelo, asociada a la métrica NCFB, está entre 0 y 20% del total de casos. Por el contrario, cuando la precisión para clasificar casos en alguna de las otras clases distintas de la clase A toma un valor BAJO, el porcentaje de clasificaciones incorrectas para esa clase, asociada a la métrica NCFB, está entre 10 y 30% del total de los casos, mientras que el porcentaje total de clasificaciones

incorrectas para la clase A, asociada a la métrica NCFA, está entre 0 y 20% del total de casos.

TABLA LXXII. DISTRIBUCIÓN DE MODELOS CON VALOR PDM (CLASE A)= BAJO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)			
		0	*10*	*20*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	*50*	0	0	10	**10**
	60	0	20	0	**20**
	70	20	10	0	**30**
	80	10	0	0	**10**
	Total	**30**	**30**	**10**	*70*

TABLA LXXIII. DISTRIBUCIÓN DE MODELOS CON VALOR PDM (CLASE B)= BAJO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)				
		50	*60*	*70*	*80*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	*0*	0	0	20	10	**30**
	10	0	20	10	0	**30**
	20	10	0	0	0	**10**
	Total	**10**	**20**	**30**	**10**	**70**

En las Tablas LXXIV y LXXV, se muestra la distribución del número de modelos en función del porcentaje de casos clasificados incorrectamente y de la precisión obtenida para cada clase del modelo. En la misma, se observa que para que el grado de precisión de un modelo sea BAJO se necesita entre 20 y 30% de casos clasificados incorrectamente en el modelo.

TABLA LXXIV. DISTRIBUCIÓN DE MODELOS CON VALOR PDM (CLASE A)= BAJO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase A incorrectas - (NCFA)			
		10	*20*	*30*	*Total*
% Clasificaciones como clase B incorrectas - (NCFB)	*0*	0	10	20	**30**
	10	10	20	0	**30**
	20	10	0	0	**10**
	Total	**20**	**30**	**20**	*70*

TABLA LXXV. DISTRIBUCIÓN DE MODELOS CON VALOR PDM (CLASE B)= BAJO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase A incorrectas - (NCFA)			
		0	*10*	*20*	*Total*
% Clasificaciones como clase B incorrectas - (NCFB)	*10*	0	10	10	**20**
	20	10	20	0	**30**
	30	20	0	0	**20**
	Total	**30**	**30**	**10**	*70*

Con esta información, se decide estudiar el comportamiento de aquellos modelos que mejor caracterizan el grado de la métrica PDM (C) con valor BAJO.

Del análisis experimental, se observa que la métrica PDM (C) instancia una mayor cantidad de modelos con valor BAJO en la precisión para la clase A, cuando el porcentaje de clasificaciones correctas para esta clase, asociada a la métrica NCVA, es menor al 20% del total de casos, mientras que el porcentaje total de clasificaciones correctas en otras clases del modelo, asociada a la métrica NCVB, está entre 60 y 70% del total de casos.

Por el contrario, cuando la precisión para clasificar casos en alguna de las otras clases distintas de la clase A toma un valor BAJO, el porcentaje de clasificaciones correctas para esa clase, asociada a la métrica NCVB, es menor al 20% del total de los casos, mientras que el porcentaje total de clasificaciones correctas para la clase A, asociada a la métrica NCVA, está entre 60 y 70% del total de casos. En ambas situaciones, la proporción de clasificaciones correctas en cada clase o grupo representa el 71% de los modelos simulados con BAJO grado de precisión de clasificación, obteniéndose un valor medio relativo para la métrica PDM (C) del 26%. En las tablas LXXVI y LXXVII, se indican aquellos modelos con mayor incidencia en la métrica PDM (C) en función del porcentaje de casos clasificados correctamente y de la precisión obtenida para cada clase o grupo del modelo.

TABLA LXXVI. MODELOS CON MAYOR INCIDENCIA SOBRE PDM (CLASE A)= BAJO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)			
		0	*10*	*20*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	*50*	0	0	10	**10**
	60	**0**	**20**	0	**20**
	70	**20**	**10**	0	**30**
	80	10	0	0	**10**
	Total	**30**	**30**	**10**	*70*

TABLA LXXVII. MODELOS CON MAYOR INCIDENCIA SOBRE PDM (CLASE B)= BAJO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)				
		50	*60*	*70*	*80*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	*0*	0	**0**	**20**	10	**30**
	10	0	**20**	**10**	0	**30**
	20	10	0	0	0	**10**
	Total	**10**	**20**	**30**	**10**	*70*

Por otra parte, para los modelos que más inciden sobre la métrica PDM (C) se observa que, cuando la precisión del modelo para clasificar casos en la clase A toma un valor BAJO, el porcentaje de clasificaciones incorrectas para la clase A, asociada a la métrica NCFA, está entre 20 y 30% del total de casos, mientras que el porcentaje total de clasificaciones incorrectas en otras clases del modelo, asociada a la métrica NCFB, es menor al 20% del total de casos. Por el contrario, cuando la precisión para clasificar casos en alguna de las otras

clases distintas de la clase A toma un valor BAJO, el porcentaje de clasificaciones incorrectas para esa clase, asociada a la métrica NCFB, está entre 20 y 30% del total de los casos, mientras que el porcentaje total de clasificaciones incorrectas para la clase A, asociada a la métrica NCFA, es menor al 20% del total de casos, tal como se muestra en las Tablas LXXVIII y LXXIX. Esta proporción de clasificaciones incorrectas representa el 80% de los modelos de explotación de información con mayor incidencia sobre la métrica en estudio.

TABLA LXXVIII. MODELOS CON MAYOR INCIDENCIA SOBRE PDM (CLASE A)= BAJO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase A incorrectas - (NCFA)			
		10	*20*	*30*	*Total*
% Clasificaciones como clase B incorrectas - (NCFB)	*0*	0	**10**	**10**	**20**
	10	0	**20**	**0**	**20**
	20	10	0	0	**10**
	Total	**10**	**30**	**10**	***50***

TABLA LXXIX. MODELOS CON MAYOR INCIDENCIA SOBRE PDM (CLASE B)= BAJO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase A incorrectas - (NCFA)			
		0	*10*	*20*	*Total*
% Clasificaciones como clase B incorrectas - (NCFB)	*10*	0	0	10	**10**
	20	**10**	**20**	0	**30**
	30	**10**	**0**	0	**10**
	Total	**20**	**20**	**10**	***50***

Del análisis realizado a los modelos que más inciden cuando la métrica PDM (C) toma un valor BAJO, se concluye que cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos es menor al 20% del total de casos, el número de casos clasificados incorrectamente en esa misma clase está entre 20 y 30% y la exactitud del modelo está entre 70 y 80%, se obtiene la mayor cantidad de modelos de explotación de información con características similares y un valor BAJO para la métrica PDM (C). Esto representa el 57% de los modelos simulados con BAJO grado de precisión para clasificar una clase, obteniéndose un valor medio relativo para la métrica PDM (C) del 23%. Sin embargo, cuando el número de casos clasificados incorrectamente en al menos una de las clases es menor al 10% del total de casos, manteniendo el número de casos clasificados correctamente en esa misma clase, el valor relativo para la métrica PDM (C) desciende a 0%, lo que representa el 14% de los modelos simulados.

d) Regla Experimental

Como regla experimental del comportamiento de la métrica analizada, se concluye que:

- El grado de precisión de un modelo es ALTO, cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos, asociado a la métrica NCVCi, está entre 20 y 70% del total de casos (equivale de 11 a 70% de casos), el número de casos clasificados incorrectamente en esa misma clase, asociado a la métrica NCFCi, está entre 0 y 10% y la exactitud del modelo es superior al 80%. En la Fig. 14 se muestra la proporción de casos necesarios clasificados correcta e incorrectamente para un modelo por cada métrica analizada. Esta regla de comportamiento cubre el 60% de los modelos simulados con alta precisión para clasificar casos a una clase.
- El grado de precisión de un modelo es MEDIO, cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos, asociado a la métrica NCVCi, está entre 10 y 40% del total de casos (equivale de 1 a 40% de casos), el número de casos clasificados incorrectamente en esa misma clase, asociado a la métrica NCFCi, está entre 10 y 20% (equivale de 1 a 20% de casos) y la exactitud del modelo está entre 70 y 90%. En la Fig. 14 se muestra la proporción de casos necesarios clasificados correcta e incorrectamente para un modelo por cada métrica analizada. Esta regla de comportamiento cubre el 75% de los modelos simulados con precisión media para clasificar casos a una clase.
- El grado de precisión de un modelo es BAJO, cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos, asociado a la métrica NCVCi, está entre 0 y 10% del total de casos, el número de casos clasificados incorrectamente en esa misma clase, asociado a la métrica NCFCi, está entre 20 y 30% (equivale de 11 a 30% de casos) y la exactitud del modelo está entre 70 y 80%. En la Fig. 14 se muestra la proporción de casos necesarios clasificados correcta e incorrectamente para un modelo por cada métrica analizada. Esta regla de comportamiento cubre el 57% de los modelos simulados con baja precisión para clasificar casos a una clase.

Cabe mencionar, que a medida que aumenta el número de clases o grupos utilizados en la construcción de un modelo de Descubrimiento de Reglas, que aplica algoritmos de inducción TDIDT, la performance del mismo tiende a decrecer [39].

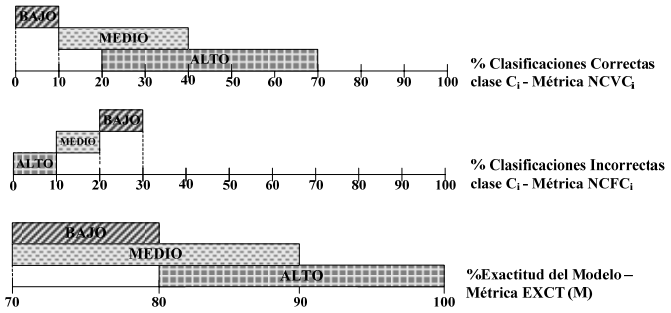

Fig. 14. Proporción de casos clasificados correctos e incorrectos según rango en métrica PDM (C)

5. Métrica de Tasa de Aciertos y Tasa de Errores del Modelo

Para estudiar la métrica de Tasa de Aciertos y Errores del Modelo para clasificar casos a una clase – TAM (C) y TEM (C), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de las mismas.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en las métricas TAM (C) y TEM (C), consideradas como métricas derivada. La lista de variables experimentales se muestra en la Tabla LXXX.

En la sección de diseño experimental (sub-sección *b*) se establece como criterio para la simulación, los valores asociados al número de casos a utilizar en la fase de prueba del

modelo del proyecto de explotación de información, ya que la tasa de aciertos y de errores de un modelo es un indicador para evaluar la calidad del mismo, luego de construido y entrenado.

TABLA LXXX. VARIABLES EXPERIMENTALES PARA LAS MÉTRICAS TAM (M) Y TEM (M)

Variable Experimental	Descripción
NCVA	Número de casos pertenecientes a la clase A que fueron correctamente clasificados por el modelo en esa misma clase.
NCVB	Número de casos pertenecientes a la clase B que fueron correctamente clasificados por el modelo en esa misma clase. Para un problema de N clases, representa todos aquellos casos que fueron clasificados correctamente en otras clases diferentes de la clase A.
NCFA	Número de casos que fueron incorrectamente clasificados por el modelo como clase A, pero que pertenecen a otra clase.
NCFB	Número de casos que pertenecen a la clase A pero que fueron incorrectamente clasificados por el modelo en otra clase. Para un problema de N clases, representa todos aquellos casos que fueron clasificados incorrectamente en otras clases diferentes de la clase A.
NCP (M)	Número de casos a utilizar para las pruebas del modelo. Para un problema de N clases, representa la cantidad de casos de prueba utilizados por cada clase C_i del modelo.
NCP (C)	Número de casos a utilizar para las pruebas de una clase o grupo del modelo.
TAM (C)	Tasa de aciertos del modelo para clasificar casos a una clase, la cual se define como: $TAM(C_i) = \frac{NCVC_i}{NCVC_i + NCFC_{\overline{i}}} = \frac{NCVC_i}{NCP(C_i)}$ donde C_i es la clase correspondiente que se está analizando y $\overline{C_i}$ son las restantes clases del modelo.
TEM (C)	Tasa de errores del modelo en la clasificación de casos a una clase, la cual se define como: $TEM(C_i) = 1 - TAM(C_i)$ donde C_i es la clase correspondiente que se está analizando.

b) Diseño Experimental

Para analizar el comportamiento de las métricas enunciadas, se utiliza un banco de pruebas simulado de 760 modelos de explotación de información de prueba, definiendo un rango de valores específico para cada una de las variables experimentales independientes y considerando las restricciones indicadas en [65] según el tamaño del proyecto. Además, se considera que la exactitud de estos modelos de prueba es alta y que se aplican a modelos de explotación de información previamente entrenados, con un alto nivel de exactitud en la clasificación de clases y un nivel alto y/o medio de precisión.

Estas consideraciones corresponden a los criterios que normalmente se toman en cuenta al momento de evaluar la calidad de los modelos construidos para un proyecto de explotación de información real. De esta manera, se generan diferentes combinaciones sujetas a análisis. Los valores simulados para las variables experimentales independientes se muestran en la Tabla LXXXI.

El resultado y comportamiento de las métricas derivadas TAM (C) y TEM (C) está asociado al número de casos de prueba clasificados correcta e incorrectamente por el modelo de explotación de información en una determinada clase o grupo, respecto del total de casos que pertenecen a esa misma clase, cuyas métricas básicas NCVA, NCVB, NCFA y NCFB, se indican en la Tabla LXXXI. A su vez, las primeras dos métricas representan la cantidad total de casos clasificados correctamente por el modelo, mientras que las dos últimas métricas representan la cantidad total de casos clasificados incorrectamente.

TABLA LXXXI. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LAS MÉTRICAS TAM (M) Y TEM

Variable Experimental	Descripción
NCVA	Número de casos pertenecientes a la clase A que fueron correctamente clasificados por el modelo en esa misma clase, con un rango de valores específico 0, 10, 20, 30, 40, 50, 60, 70, 80 y 90. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en el modelo.
NCVB	Número de casos pertenecientes a la clase B que fueron correctamente clasificados por el modelo en esa misma clase, con un rango de valores específico 0, 10, 20, 30, 40, 50, 60, 70, 80 y 90. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en el modelo.
NCFA	Número de casos que fueron incorrectamente clasificados por el modelo como clase A, pero que pertenecen a otra clase, con un rango de valores específico 0, 10, 20 y 30. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en el modelo.
NCFB	Número de casos que pertenecen a la clase A pero que fueron incorrectamente clasificados por el modelo en otra clase, con un rango de valores específico 0, 10, 20 y 30. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en el modelo.
NCP (M)	Número de casos a utilizar para la prueba del modelo, con un rango de valores específico 300.000, 600.000, 900.000, 1.200.000, 1.500.000, 1.800.000, 2.100.000, 2.400.000, 2.700.000 y 3.000.000, según lo predefinido para un proyecto de tamaño pequeño.

Por otra parte, se tiene como restricción que la cantidad de casos clasificados correcta e incorrectamente debe cubrir la totalidad de los casos utilizados para la prueba del modelo construido.

$$NCP(M) = NCP(C_A) + NCP(C_B)$$

donde:

- $NCP(C_A) = NCVA + NCFB$ es el número de casos de prueba para la clase A.
- $NCP(C_B) = NCVB + NCFA$ es el número de casos de prueba para la clase B.

Al tratarse de métricas básicas sometidas a un proceso de simulación, no se les puede asignar un número a priori de casos a utilizar. Por tal motivo, se decide establecer un rango de valores porcentuales con intervalos de a 10%, definido entre 0 y 90% para las clasificaciones correctas y entre 0 y 30% para las clasificaciones incorrectas, respecto del número total de casos del modelo, tal como se indica en las Tablas LXXXII y LXXXIII. Estos valores, a su vez, comprenden a los modelos de explotación de información cuya métrica de exactitud es superior al 70% (sub-sección *3*).

A partir de los rangos de valores definidos, se analiza el comportamiento general de las métricas TAM (C) y TEM (C).

Como se indicó anteriormente, existe la restricción entre estas métricas respecto del total de casos de prueba del modelo. Por consiguiente, se necesitan sólo aquellas configuraciones de las métricas NCVA, NCVB, NCFA y NCFB, cuya suma represente el 100% de los casos. De este análisis se obtienen 76 configuraciones posibles aplicables a un modelo.

TABLA LXXXII. RANGOS Y VALORES RELACIONADOS A LAS MÉTRICAS NCVA Y NCVB PARA CLASIFICACIONES CORRECTAS

Métrica	Descripción	Valores de simulación
NCVA NCVB	Ningún caso clasificado correctamente	0
	De 1 a 10% de casos clasificados correctamente	10
	De 11 a 20% de casos clasificados correctamente	20
	De 21 a 30% de casos clasificados correctamente	30
	De 31 a 40% de casos clasificados correctamente	40
	De 41 a 50% de casos clasificados correctamente	50
	De 51 a 60% de casos clasificados correctamente	60
	De 61 a 70% de casos clasificados correctamente	70
	De 71 a 80% de casos clasificados correctamente	80
	De 81 a 90% de casos clasificados correctamente	90

TABLA LXXXIII. RANGOS Y VALORES RELACIONADOS A LAS MÉTRICAS NCFA Y NCFB PARA CLASIFICACIONES INCORRECTAS

Métrica	Descripción	Valores de simulación
NCFA NCFB	Ningún caso clasificado incorrectamente	0
	De 1 a 10% de casos clasificados incorrectamente	10
	De 11 a 20% de casos clasificados incorrectamente	20
	De 21 a 30% de casos clasificados incorrectamente	30

En virtud que no existe un criterio universal que defina qué indicador se debe adoptar para considerar un modelo con ALTA o BAJA tasa de aciertos y errores en la clasificación de casos a una clase en un proyecto de explotación de información, en las Tablas LXXXIV y LXXXV se indican los valores y rangos contemplados para un proyecto de explotación de información para PyMEs, considerados a partir de la propia experiencia.

TABLA LXXXIV. RANGOS Y VALORES RELACIONADOS A LA MÉTRICA TAM (M)

Rango	Descripción
Bajo	Menos de 70% de casos de prueba clasificados correctamente en su clase
Alto	Más de 70% de casos de prueba clasificados correctamente en su clase

TABLA LXXXV. RANGOS Y VALORES RELACIONADOS A LA MÉTRICA TEM (M)

Rango	Descripción
Bajo	Menos de 30% de casos de prueba clasificados incorrectamente en su clase
Alto	Más de 30% de casos de prueba clasificados incorrectamente en su clase

A partir de la asignación de los valores a cada variable independiente, se generan los datos de los modelos de explotación de información de prueba. De esta manera se busca obtener los resultados y el comportamiento de las métricas TAM (C) y TEM (C) (variables experimentales dependientes) en función de la variación de estas variables.

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de las métricas TAM (C) y TEM (C), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que la métrica derivada TAM (C) toma un valor ALTO, considerándose como parámetro para este análisis un valor relativo superior al 70%, cuando el porcentaje de casos de prueba clasificados correctamente para las clases A y B, asociado a las métricas NCVA y NCVB, está entre 10 y 90% del total de casos de prueba utilizados en el modelo. En la Tabla LXXXVI, se muestra la distribución del número de modelos en función del porcentaje de casos clasificados correctamente. En esta tabla, se observa que para que la tasa de aciertos de un modelo sea ALTO y la tasa de errores sea BAJA, se necesita que el mismo tenga una exactitud mínima del 80% de casos de prueba clasificados correctamente en su clase o grupo.

Por otra parte, el porcentaje de clasificaciones incorrectas en cada una de las clases, asociadas a las métricas NCFB y NCFA, está entre 0 y 20% del total de casos del modelo. En la Tabla LXXXVII, se muestra la distribución del número de modelos en función del porcentaje de casos de prueba clasificados incorrectamente en cada clase. En la misma, se observa que para que la tasa de aciertos de un modelo sea ALTO y la tasa de errores sea BAJA, se necesita un máximo de 20% de casos de prueba clasificados incorrectamente en otras clases.

TABLA LXXXVII. DISTRIBUCIÓN DE MODELOS CON VALOR TAM (C)=ALTO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase B incorrectas - (NCFB)			
		0	***10***	***20***	***Total***
% Clasificaciones como clase A incorrectas - (NCFA)	***10***	90	60	30	**180**
	20	60	30	0	**90**
	30	30	0	0	**30**
	Total	**180**	**90**	**30**	***300***

Con esta información, se decide estudiar el comportamiento de aquellos modelos que mejor caracterizan el grado de la métrica TAM (C) con valor ALTO y de la métrica TEM (C) con valor BAJO.

Del análisis experimental, se observa que la métrica TAM (C) instancia una mayor cantidad de modelos con valor ALTO, cuando el porcentaje de casos clasificados correctamente para las clases A y B, asociado a las métricas NCVA y NCVB, está entre 20 y 70% del total de casos del modelo. Esta proporción de clasificaciones correctas en cada clase o grupo representa el 73% de los modelos simulados con ALTA tasa de aciertos y BAJA tasa de errores, obteniéndose un valor medio relativo para la métrica TAM (C) del 91% y para la métrica TEM (C) del 9%. En la Tabla LXXXVIII, se indican aquellos modelos con mayor incidencia en las métricas TAM (C) y TEM (C) en función del porcentaje de casos clasificados correctamente en cada clase o grupo. Por otra parte, para los modelos que más inciden sobre las métricas TAM (C) y TEM (C), se observa que en mayor proporción el porcentaje de casos de prueba clasificados incorrectamente, asociados a las métricas NCFB y NCFA, está entre 0 y 10% del total de casos del modelo, tal como se muestra en la Tabla LXXXIX.

TABLA LXXXVI. DISTRIBUCIÓN DE MODELOS CON VALOR TAM (C)=ALTO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)									
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	*10*	0	0	0	0	0	0	10	10	10	**30**
	20	0	0	0	0	0	10	10	10	0	**30**
	30	0	0	0	0	20	20	10	0	0	**50**
	40	0	0	0	10	20	10	0	0	0	**40**
	50	0	0	20	20	10	0	0	0	0	**50**
	60	0	10	20	10	0	0	0	0	0	**40**
	70	10	10	10	0	0	0	0	0	0	**30**
	80	10	10	0	0	0	0	0	0	0	**20**
	90	10	0	0	0	0	0	0	0	0	**10**
	Total	**30**	**30**	**50**	**40**	**50**	**40**	**30**	**20**	**10**	*300*

TABLA LXXXVIII. MODELOS CON MAYOR INCIDENCIA SOBRE TAM (C)=ALTO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)									
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*Total*
% Clasificaciones como clase B correctas - (NCVB)	*10*	0	**0**	**0**	**0**	**0**	**0**	10	10	10	**30**
	20	0	**0**	**0**	**0**	**0**	**10**	**10**	10	0	**30**
	30	0	**0**	**0**	**0**	**20**	**20**	**10**	0	0	**50**
	40	0	**0**	**0**	**10**	**20**	**10**	**0**	0	0	**40**
	50	0	**0**	**20**	**20**	**10**	**0**	**0**	0	0	**50**
	60	0	**10**	**20**	**10**	**0**	**0**	**0**	0	0	**40**
	70	10	**10**	**10**	**0**	**0**	**0**	**0**	0	0	**30**
	80	10	10	0	0	0	0	0	0	0	**20**
	90	10	0	0	0	0	0	0	0	0	**10**
	Total	**30**	**30**	**50**	**40**	**50**	**40**	**30**	**20**	**10**	*300*

TABLA LXXXIX. MODELOS CON MAYOR INCIDENCIA SOBRE TAM (C)= ALTO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase B incorrectas - (NCFB)			
		0	*10*	*20*	*Total*
% Clasificaciones como clase A incorrectas - (NCFA)	*10*	**50**	**50**	20	**120**
	20	**50**	**30**	0	**80**
	30	20	0	0	**20**
	Total	**120**	**80**	**30**	*220*

Esta proporción de clasificaciones incorrectas representa el 82% de los modelos de explotación de información con mayor incidencia sobre las métricas en estudio, mientras que el 18% restante presenta un comportamiento extrapolado respecto a los casos clasificados incorrectamente para cada clase.

Del análisis realizado a los modelos que más inciden cuando la métrica TAM (C) toma un valor ALTO y la métrica TEM (C) toma un valor BAJO, se concluye que cuando en una clase o grupo cualquiera, el número de casos de prueba clasificados correctamente está entre 20 y 70% del total de casos, el número de casos de prueba de esa misma clase pero clasificados incorrectamente en otras clases está entre 0 y 10% y la exactitud del modelo es superior al 80%, se obtiene la mayor cantidad de modelos de explotación de información con características similares, un valor ALTO para la métrica TAM (C) y un valor BAJO para la métrica TEM (C). Esto representa el 60% de los modelos simulados con ALTA tasa de aciertos de casos y BAJA tasa de errores, obteniéndose un valor medio relativo para la métrica TAM (C) del 92% y para la métrica TEM (C) del 8%.

Sin embargo, cuando el número de casos de prueba clasificados incorrectamente impacta sólo en una de las clases (una de las métricas toma valor 20% y la otra 0%), manteniendo el número de casos clasificados correctamente, el valor medio relativo para la métrica TAM (C) desciende al 73% mientras que la métrica TEM (C) asciende al 27%, lo que representa el 13% de los modelos simulados.

En el segundo análisis experimental realizado sobre la métrica TAM (C) y TEM (C), se observa que cuando la primera toma un valor BAJO y la segunda un valor ALTO, se presentan dos situaciones que afectan la tasa de aciertos y errores de cada clase en forma individual dependiendo del número de casos clasificados correcta e incorrectamente en cada una de las clases o grupos. En la primera situación, se observa que la tasa de aciertos del modelo para los casos de prueba de la clase A toma un valor BAJO (la tasa de aciertos total de las otras clases del modelo toma un valor alto), cuando el porcentaje de casos de prueba clasificados correctamente para la clase A, asociado a la métrica NCVA, está entre 0 y

60% del total de casos utilizados en el modelo. A su vez, el porcentaje de casos de prueba clasificados correctamente en otras clases, correspondiente a la clase B y asociado a la métrica NCVB, está entre 10 y 80%.

En este caso, la tasa de errores para los casos de prueba pertenecientes a la clase A toma un valor ALTO. Por el contrario, cuando la tasa de aciertos de las otras clases distintas de la clase A toma un valor BAJO, la tasa de aciertos y de errores en la clase A es alta y baja, respectivamente. En este caso, el porcentaje de clasificaciones correctas para otras clases, asociada a la métrica NCVB, está entre 0 y 60% del total de los casos de prueba utilizados en el modelo, mientras que el porcentaje total de clasificaciones correctas para la clase A, asociada a la métrica NCVA, está entre 10 y 80%.

En las Tablas XC y XCI, se muestra la distribución del número de modelos en función del porcentaje de casos de prueba clasificados correctamente y de la tasa de aciertos obtenida para cada clase del modelo. En las mismas, se observa que para que la tasa de aciertos de un modelo sea BAJA y la tasa de errores sea ALTA, se necesita que el mismo tenga una exactitud entre 70 y 90% de casos de prueba clasificados correctamente en su clase o grupo.

A los fines del análisis del comportamiento de las métricas y de las implicancias para el proyecto de explotación de información, se asume que cuando la tasa de aciertos de alguna de las clases del modelo construido toma un valor BAJO, la tasa de aciertos global del modelo es BAJA.

TABLA XC. DISTRIBUCIÓN DE MODELOS CON VALOR TAM (CLASE)=BAJO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)							
		0	***10***	***20***	***30***	***40***	***50***	***60***	***Total***
% Clasificaciones como clase B correctas - (NCVB)	***10***	0	0	0	0	0	0	10	**10**
	20	0	0	0	0	0	10	0	**10**
	30	0	0	0	0	20	0	0	**20**
	40	0	0	0	20	10	0	0	**30**
	50	0	0	30	10	0	0	0	**40**
	60	0	30	20	0	0	0	0	**50**
	70	20	20	10	0	0	0	0	**50**
	80	10	10	0	0	0	0	0	**20**
	Total	**30**	**60**	**60**	**30**	**30**	**10**	**10**	***230***

TABLA XCI. DISTRIBUCIÓN DE MODELOS CON VALOR TAM (CLASE B)= BAJO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)								
		10	***20***	***30***	***40***	***50***	***60***	***70***	***80***	***Total***
% Clasificaciones como clase B correctas - (NCVB)	***0***	0	0	0	0	0	0	20	10	**30**
	10	0	0	0	0	0	30	20	10	**60**
	20	0	0	0	0	30	20	10	0	**60**
	30	0	0	0	20	10	0	0	0	**30**
	40	0	0	20	10	0	0	0	0	**30**
	50	0	10	0	0	0	0	0	0	**10**
	60	10	0	0	0	0	0	0	0	**10**
	Total	**10**	**10**	**20**	**30**	**40**	**50**	**50**	**20**	***230***

Por otra parte, cuando la tasa de aciertos del modelo para los casos de prueba de la clase A toma un valor BAJO, el porcentaje de casos de esa misma clase clasificado incorrectamente en otras clases, asociada a la métrica NCFB, está entre 10 y 30% del total de casos.

A su vez, el porcentaje total de casos de prueba clasificados incorrectamente por el modelo en la clase A, asociado a la métrica NCFA, está entre 0 y 20% del total de casos.

Por el contrario, cuando la tasa de aciertos para los casos de prueba de las otras clases distintas de la clase A toma un valor BAJO, el porcentaje de casos de esa misma clase clasificado incorrectamente en otras clases, asociada a la métrica NCFA, está entre 10 y 30% del total de los casos. A su vez, que el porcentaje total de casos de prueba clasificados incorrectamente por el modelo en esa misma clase, asociado a la métrica NCFB, está entre 0 y 20% del total de casos. En las Tablas XCII y XCIII, se muestra la distribución del número de modelos en función del porcentaje de casos de prueba clasificados incorrectamente en cada clase y de la tasa de aciertos obtenida para cada clase del modelo. En la misma, se observa que para que la tasa de aciertos de un modelo sea BAJA y la tasa de errores sea ALTA, se necesita un máximo de 30% de casos de prueba clasificados incorrectamente en otras clases.

Con esta información, se decide estudiar el comportamiento de aquellos modelos que mejor caracterizan el grado de la métrica TAM (C) con valor BAJO y de la métrica TEM (C) con valor ALTO.

TABLA XCII. DISTRIBUCIÓN DE MODELOS CON VALOR TAM (CLASE)=BAJO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase B incorrectas - (NCFB)			
		0	***10***	***20***	***Total***
% Clasificaciones como clase A incorrectas - (NCFA)	***10***	20	40	60	**120**
	20	30	50	0	**80**
	30	30	0	0	**30**
	Total	**80**	**90**	**60**	***230***

TABLA XCIII. DISTRIBUCIÓN DE MODELOS CON VALOR TAM (CLASE)=BAJO RESPECTO DE LAS MÉTRICAS NCFA Y NCFB

		% Clasificaciones como clase B incorrectas - (NCFB)			
		10	***20***	***30***	***Total***
% Clasificaciones como clase A incorrectas - (NCFA)	***0***	20	30	30	**80**
	10	40	50	0	**90**
	20	60	0	0	**60**
	Total	**120**	**80**	**30**	***230***

Del análisis experimental, se observa que la métrica TAM (C) instancia una mayor cantidad de modelos con valor BAJO, cuando el porcentaje de clasificaciones correctas para la clase A, asociado a la métrica NCVA, está entre 0 y 30% del total de casos del modelo. A su vez, el porcentaje casos clasificados correctamente en otras clases del modelo, asociado a la métrica NCVB, está entre 40 y 70% del total de casos.

Por el contrario, cuando la tasa de aciertos de las otras clases distintas de la clase A toma un valor BAJO, el porcentaje de clasificaciones correctas para esa clase, asociada a la métrica NCVB, está entre 0 y 30% del total de los casos, mientras que el porcentaje total de clasificaciones correctas para la clase A, asociada a la métrica NCVA, está entre 40 y 70% del total de casos. En ambas situaciones, la proporción de clasificaciones correctas en cada clase o grupo representa el 70% de los modelos simulados con BAJA tasa de aciertos y ALTA tasa de errores, obteniéndose un valor medio relativo para la métrica TAM (C) del 44% y para la métrica TEM (C) del 56%. En las Tablas XCIV y XCV, se indican aquellos modelos con mayor incidencia en las métricas TAM (C) y TEM (C) en función del porcentaje de casos clasificados correctamente en cada clase o grupo.

TABLA XCIV. MODELOS CON MAYOR INCIDENCIA SOBRE TAM (CLASE A)=BAJO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)							
		0	***10***	***20***	***30***	***40***	***50***	***60***	***Total***
% Clasificaciones como clase B correctas - (NCVB)	***10***	0	0	0	0	0	0	10	**10**
	20	0	0	0	0	0	10	0	**10**
	30	0	0	0	0	20	0	0	**20**
	40	**0**	**0**	**0**	**20**	10	0	0	**30**
	50	**0**	**0**	**30**	**10**	0	0	0	**40**
	60	**0**	**30**	**20**	**0**	0	0	0	**50**
	70	**20**	**20**	**10**	**0**	0	0	0	**50**
	80	10	10	0	0	0	0	0	**20**
	Total	**30**	**60**	**60**	**30**	**30**	**10**	**10**	***230***

TABLA XCV. MODELOS CON MAYOR INCIDENCIA SOBRE TAM (CLASE B)=BAJO RESPECTO DE LAS MÉTRICAS NCVA Y NCVB

		% Clasificaciones como clase A correctas - (NCVA)								
		10	***20***	***30***	***40***	***50***	***60***	***70***	***80***	***Total***
% Clasificaciones como clase B correctas - (NCVB)	***0***	0	0	0	**0**	**0**	**0**	**20**	10	**30**
	10	0	0	0	**0**	**0**	**30**	**20**	10	**60**
	20	0	0	0	**0**	**30**	**20**	**10**	0	**60**
	30	0	0	0	**20**	**10**	**0**	**0**	0	**30**
	40	0	0	20	10	0	0	0	0	**30**
	50	0	10	0	0	0	0	0	0	**10**
	60	10	0	0	0	0	0	0	0	**10**
	Total	**10**	**10**	**20**	**30**	**40**	**50**	**50**	**20**	***230***

Por otra parte, para los modelos que más inciden sobre las métricas TAM (C) y TEM (C), se observa que cuando la tasa de aciertos del modelo para los casos de prueba de la clase A toma un valor BAJO, el porcentaje de casos de prueba de esa misma clase pero clasificados incorrectamente en otras clases, asociado a la métrica NCFB, está entre 10 y 20% del total de casos del modelo, mientras que el porcentaje total de clasificaciones incorrectas en esa misma clase, asociada a la métrica NCFA, es menor al 20% del total de casos. Por el contrario, cuando la tasa de aciertos de las otras clases distintas de la clase A toma un valor BAJO, el porcentaje de casos de esa misma clase clasificado incorrectamente en otras clases, asociado a la métrica NCFA, está entre 10 y 20% del total de los casos. A su vez, el porcentaje total de casos de prueba clasificados incorrectamente por el modelo en esa misma clase, asociado a la métrica NCFB, es menor al 20% del total de casos, tal como se muestran en las Tablas XCVI y XCVII. Esta proporción de clasificaciones incorrectas representa el 63% de los modelos de explotación de información con mayor incidencia sobre las métricas en estudio, mientras que el 37% restante presenta un comportamiento extrapolado respecto a los casos clasificados incorrectamente para cada clase.

TABLA XCVI. MODELOS CON MAYOR INCIDENCIA SOBRE TAM (CLASE A)=BAJO RESPECTO DE LAS MÉTRICAS NCFB Y NCFA

		% Clasificaciones como clase B incorrectas - (NCFB)			
		10	***20***	***30***	***Total***
% Clasificaciones como clase A incorrectas - (NCFA)	***0***	**10**	**30**	30	**70**
	10	**20**	**40**	0	**60**
	20	30	0	0	**30**
	Total	**60**	**70**	**30**	***160***

TABLA XCVII. MODELOS CON MAYOR INCIDENCIA SOBRE TAM (CLASE B)=BAJO RESPECTO DE LAS MÉTRICAS NCFB Y NCFA

		% Clasificaciones como clase B incorrectas - (NCFB)			
		0	***10***	***20***	***Total***
% Clasificaciones como clase A incorrectas - (NCFA)	***10***	**10**	**20**	30	**60**
	20	**30**	**40**	0	**70**
	30	30	0	0	**30**
	Total	**70**	**60**	**30**	***160***

Del análisis realizado a los modelos que más inciden cuando la métrica TAM (C) toma un valor BAJO y la métrica TEM (C) toma un valor ALTO, se concluye que cuando en una clase o grupo cualquiera, el número de casos de prueba clasificados correctamente está entre 0 y 30% del total de casos, el número de casos de prueba de esa misma clase pero clasificados incorrectamente en otras clases está entre 10 y 20% y la exactitud del modelo está entre 70 y 90%, se obtiene la mayor cantidad de modelos de explotación de información con características similares, un valor BAJO para la métrica TAM (C) y un valor ALTO para la métrica TEM (C). Esto representa aproximadamente el 44% de los modelos simulados con BAJA tasa de aciertos de casos y ALTA tasa de errores, obteniéndose un valor medio relativo para la métrica TAM (C) del 47% y para la métrica TEM (C) del 53%.

d) Regla Experimental

Como regla experimental del comportamiento de la métrica analizada, se concluye que:

- La tasa de aciertos de un modelo es ALTA y la tasa de errores es BAJA, cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos, asociado a la métrica NCVCi, está entre 20 y 70% del total de casos (equivale de 11 a 70% de casos), el número de casos de esa misma clase pero clasificados incorrectamente en otras clases, asociado a la métrica $NCF\,\overline{C_i}$, está entre 0 y 10% y la exactitud del modelo es superior al 80%. En las Fig. 15 y 16 se muestra la

proporción de casos necesarios clasificados correcta e incorrectamente para un modelo para las métricas analizadas. Esta regla de comportamiento cubre el 60% de los modelos simulados con alta tasa de aciertos y baja tasa de errores para clasificar nuevos casos a una clase.

- La tasa de aciertos de un modelo es BAJA y la tasa de errores es ALTA, cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos, asociado a la métrica NCVCi, está entre 0 y 30% del total de casos, el número de casos de esa misma clase pero clasificados incorrectamente en otras clases, asociado a la métrica NCF $\overline{C_i}$, está entre 10 y 20% (equivale de 1 a 20% de casos) y la exactitud del modelo está entre 70 y 90%. En las Fig. 15 y 16 se muestra la proporción de casos necesarios clasificados correcta e incorrectamente para un modelo para las métricas analizadas. Esta regla de comportamiento cubre el 65% de los modelos simulados con baja tasa de aciertos y alta tasa de errores para clasificar nuevos casos a una clase.

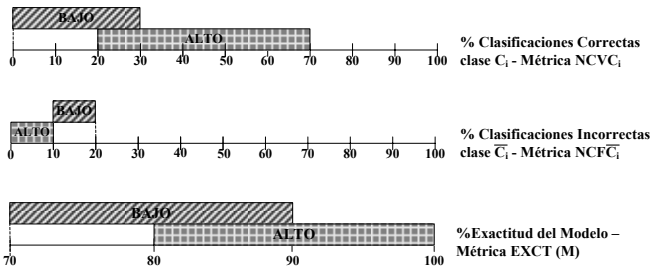

Fig. 15. Proporción de casos clasificados correctos e incorrectos según rango en métrica TAM (C)

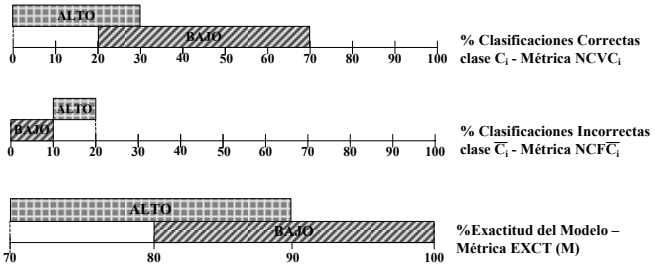

Fig. 16. Proporción de casos clasificados correctos e incorrectos según rango en métrica TEM (C)

Cabe mencionar, que a medida que aumenta el número de clases o grupos utilizados en la construcción de un modelo de Descubrimiento de Reglas, que aplica algoritmos de inducción TDIDT, la performance del mismo tiende a decrecer [39].

6. *Métrica de Cobertura de una Regla*

Para estudiar la métrica de Cobertura de una regla de comportamiento asociada a una clase o grupo del modelo – COBER (R), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de la misma.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en la métrica COBER (R), considerada como métrica derivada. La lista de variables experimentales se muestra en la Tabla XCVIII.

TABLA XCVIII. VARIABLES EXPERIMENTALES PARA LA MÉTRICA COBER (R)

Variable Experimental	Descripción
NCE (C)	Número de casos a utilizar por cada clase para en el entrenamiento del modelo. Para un problema de N clases, representa la cantidad de casos de entrenamiento utilizados por cada clase C_i del modelo.
NCCNS (R)	Número de casos que satisfacen la aplicación de la regla de comportamiento o de pertenencia de una clase o grupo.
NRGL (C)	Número de reglas descubiertas por del modelo por cada clase o grupo.
COBER (R)	Cobertura o soporte de una regla de pertenencia a una clase o grupo, la cual se define como: $\text{COBER(R)} = \frac{\text{NCCNS(R)}}{\text{NCE(C)}}$

b) Diseño Experimental

Para analizar el comportamiento de la métrica enunciada, se utiliza un banco de pruebas simulado de 500 modelos de explotación de información con la información obtenida para una clase o grupo de todo el conjunto, luego de aplicar un algoritmo de inducción TDIDT [68] al modelo construido. Para ello, se define un rango de valores específico para cada una de las variables experimentales independientes, considerando las restricciones indicadas en [65] según el tamaño del proyecto, generando diferentes combinaciones sujetas a análisis. Cabe aclarar, que las conclusiones obtenidas de analizar la cobertura de las reglas asociadas a una clase, sirven para cualquiera de las clases o grupos que tenga el modelo construido. Los valores simulados para las variables experimentales independientes se muestran en la Tabla XCIX.

TABLA XCIX. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LA MÉTRICA COBER (R)

Variable Experimental	Descripción
NCE (C)	Número de casos a utilizar para una clase o grupo del modelo durante su entrenamiento, con un rango de valores específico 1.000, 2.000, 3.000, 4.000, 5.000, 6.000, 7.000, 8.000, 9.000 y 10.000.
NCCNS (R)	Número de casos que satisfacen la aplicación de la regla, con un rango de valores específico 10, 20, 30, 40, 50, 60, 70, 80, 90 y 100. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en la clase o grupo.
NRGL (C)	Número de reglas descubiertas por del modelo para la clase o grupo, con un rango de valores específico 2, 5, 10, 15 y 20.

El resultado y comportamiento de la métrica derivada COBER (R) está asociado al número de casos de entrenamiento que son seleccionados por cada regla de pertenencia obtenida para una clase o grupo, respecto del total de casos que pertenecen a esa misma clase, luego de aplicar un modelo de Descubrimiento de Reglas, cuyas métricas básicas NCCNS (R) y NCE (C), se indican en la Tabla XCIX.

En el caso de la métrica NCCNS (R), al estar sometida a un proceso de simulación, no se le puede asignar un número a priori de casos a utilizar. Por tal motivo, se decide establecer un rango de valores porcentuales con intervalos de a 10%, definido entre 10 y 100% que represente el número de casos seleccionados por la regla, respecto del número total de casos que contiene la clase del modelo. A partir de los rangos de valores definidos, se analiza el comportamiento general de la métrica COBER (R).

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de la métrica COBER (R), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que el valor de la métrica COBER (R) aumenta cuando el porcentaje de casos que satisfacen la aplicación de la regla de pertenencia a una clase o grupo, asociado a la métrica NCCNS (R), toma valores entre 10% y 100% del total de casos de entrenamiento de la clase. Por el contrario, cuando el porcentaje de casos correspondiente a la métrica NCCNS (R) disminuye, manteniendo la misma cantidad de casos en la clase o grupo, el valor de la métrica COBER (R) también disminuye. En la Fig. 17 se muestra la proporción de casos seleccionados por una regla en función de la cobertura obtenida para esta.

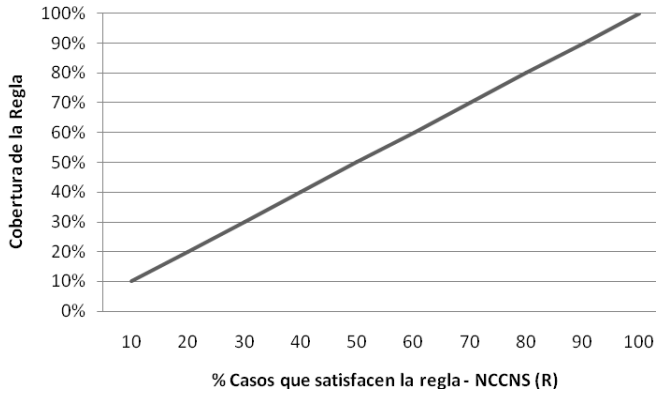

Fig. 17. Proporción de casos cubiertos por una regla para la métrica COBER (R)

La utilidad de la métrica de cobertura de una regla en un proyecto de explotación de información que utiliza un modelo de Descubrimiento de Reglas, radica en que aquellas reglas de pertenencia o asociación a una clase o grupo con mayor valor de cobertura COBER (R), son las más representativas y útiles del modelo para definir el comportamiento de esa clase o grupo. Por consiguiente, representan las reglas que mayor credibilidad e interés le otorgan al modelo construido para clasificar nuevos casos a la clase.

En virtud que no existe un criterio universal que defina qué indicador se debe adoptar como valor mínimo para la cobertura de una regla en un proyecto de explotación de información para PyMEs, se propone utilizar como valor de umbral, y a partir de la propia experiencia, el número de reglas de pertenencia obtenidas para la clase, asociada a la métrica NRGL (C), tal como se indica en la siguiente condición:

$$\text{COBER}(\text{R}) > \frac{100}{\text{NRGL}(\text{C})}$$

En un segundo análisis experimental, se observa que cuando el número de reglas que definen la pertenencia a una clase o grupo, asociado a la métrica NRGL (C), toma valores 2, 5, 10, 15 y 20, el valor de umbral mínimo que se obtiene es (expresado en porcentaje) 50%, 20%, 10%, 6.67% y 5%, respectivamente. Esto indica que, si número de reglas que describen la pertenencia de la clase es alto, el valor mínimo de umbral para evaluar la cobertura de cada regla disminuye, lo cual implica que las reglas pierdan representatividad en la cantidad de casos que cubren. Además, a medida que aumenta el número de reglas que definen la pertenencia a una clase o grupo, la performance del modelo de Descubrimiento de Reglas tiende a decrecer hasta un determinado número de reglas [39].

Cabe aclarar, que cuando existe sólo una regla de pertenencia a una clase, independientemente del valor que tome su métrica de cobertura, se considera la regla que mejor describe el comportamiento de esa clase.

d) Regla Experimental

Como regla experimental del comportamiento de la métrica COBER (R) analizada, se concluye que la misma aumenta su valor cuando se incrementa el número de casos a los que se le puede aplicar una regla de pertenencia a una clase o grupo, asociada a la métrica NCCNS (R), respecto del total de casos de esa misma clase. Mientras que, cuando el número de casos a los que se puede aplicar la regla disminuye, también lo hace la métrica de cobertura de la regla.

Por otra parte, si el número de reglas que definen la pertenencia de una clase o grupo es bajo, el valor mínimo de umbral es alto y mayor cantidad de casos debe cubrir la regla para que la misma sea representativa para el comportamiento de esa clase o grupo en el modelo de explotación de información. Por el contrario, si el número de reglas es alto, el valor mínimo de umbral es bajo y se necesita cubrir una menor cantidad de casos con la regla. No obstante, cuanto mayor es el número de reglas, más complejo resulta obtener un comportamiento determinado de los casos asociados a la clase. En consecuencia, siempre se busca que la cantidad de reglas que definen el comportamiento de una clase o grupo sea la más baja posible.

El valor de umbral propuesto para la métrica de cobertura puede ser refinado si se desea obtener mayor restricción del número de casos que debe cubrir una regla. No obstante, se dispone de un primer indicador de referencia para considerar como cantidad mínima de casos que debe cubrir una regla para ser considerada relevante y de interés para el proyecto de explotación de información.

7. Métrica de Precisión de una Regla

Para estudiar la métrica de Precisión de una regla de comportamiento asociada a una clase o grupo del modelo – PRCR (R), se toma en consideración aquellas variables experimentales (sub-sección *b*) que influyen en el comportamiento de la misma.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en la métrica PRCR (R), considerada como métrica derivada. La lista de variables experimentales se muestra en la Tabla C.

TABLA C. VARIABLES EXPERIMENTALES PARA LA MÉTRICA PRCR (R)

Variable Experimental	Descripción
NCCNS (R)	Número de casos que satisfacen la aplicación de la regla de comportamiento o de pertenencia de una clase o grupo.
NCPRC (R)	Número de casos que satisfacen la precondición de la regla, independientemente de la clase o grupo a la que pertenece.
PRCR (R)	Precisión o confianza de una regla de asociación a una clase, la cual se define como: $\text{PRCR}(\text{R}) = \frac{\text{NCCNS}(\text{R})}{\text{NCPRC}(\text{R})}$

b) Diseño Experimental

Para analizar el comportamiento de la métrica enunciada, se utiliza un banco de pruebas simulado de 100 reglas de asociación con la información obtenida para una clase o grupo de todo el conjunto, luego de aplicar un algoritmo de inducción TDIDT [68] al modelo de explotación de información construido. Para ello, se define un rango de valores específico para cada una de las variables experimentales independientes, considerando las restricciones indicadas en [65] según el tamaño del proyecto, generando diferentes combinaciones sujetas a análisis.

Cabe aclarar, que las conclusiones obtenidas de analizar la precisión de las reglas asociadas a una clase, sirven para cualquiera de las clases o grupos que tenga el modelo construido. Los valores simulados para las variables experimentales independientes se muestran en la Tabla CI.

TABLA CI. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LA MÉTRICA PRCR (R)

Variable Experimental	Descripción
NCPRC (R)	Número de casos que satisfacen la precondición de la regla, con un rango de valores específico 1.000, 2.000, 3.000, 4.000, 5.000, 6.000, 7.000, 8.000, 9.000 y 10.000.
NCCNS (R)	Número de casos que satisfacen la aplicación de la regla, con un rango de valores específico 10, 20, 30, 40, 50, 60, 70, 80, 90 y 100. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en la clase o grupo.

El resultado y comportamiento de la métrica derivada PRCR (R) se obtiene luego de aplicar un modelo de Descubrimiento de Reglas al modelo construido. Este comportamiento está asociado al número de casos de entrenamiento que son seleccionados por cada regla de pertenencia obtenida para una clase o grupo específico, respecto del total de casos que cumplen sólo con la precondición o antecedente de esa misma regla, y cuyas métricas básicas NCPRC (R) y NCCNS (R), se indican en la Tabla CI.

En el caso de la métrica NCCNS (R), al estar sometida a un proceso de simulación, no se le puede asignar un número a priori de casos a utilizar. Por tal motivo, se decide establecer un rango de valores porcentuales con intervalos de a 10%, definido entre 10 y 100% que represente el número de casos seleccionados por la regla, respecto del número total de casos que contiene la clase del modelo. A partir de los rangos de valores definidos, se analiza el comportamiento general de la métrica PRCR (R).

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de la métrica PRCR (R), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que el valor de la métrica PRCR (R) aumenta cuando el porcentaje de casos a los que se le puede aplicar una regla de pertenencia a una clase o grupo, asociado a la métrica NCCNS (R), toma valores entre 10% y 100% respecto del total de casos que sólo satisfacen la precondición o antecedente de la misma regla, asociada a la métrica NCPRC (R). Por el contrario, cuando el porcentaje de casos correspondiente a la métrica NCCNS (R) disminuye, manteniendo la misma cantidad de casos seleccionados por la precondición, el valor de la métrica PRCR (R) también disminuye. En la Fig. 18 se muestra la proporción de casos seleccionados por una regla en función de la precisión obtenida para esta.

La utilidad de la métrica de precisión de una regla en un proyecto de explotación de información que utiliza un modelo de Descubrimiento de Reglas, está en que aquellas reglas de pertenencia o asociación a una clase o grupo con mayor precisión PRCR (R), son las que mayor interés y confianza generan en el proyecto para descubrir conocimiento asociado a los casos de entrenamiento, y de aplicación a nuevos casos, cuando se aplican dichas reglas.

Sin embargo, no existe un criterio universal que defina qué indicador se debe adoptar como valor mínimo para la precisión de una regla en un proyecto de explotación de información para PyMEs. A partir de la propia experiencia, se considera que una regla de pertenencia o asociación a una clase o grupo, cuya métrica de precisión PRCR (R) toma un valor mínimo de 60-70% resulta confiable, ya que indica que toda vez que la regla se aplica en el proyecto, este porcentaje de veces se puede asociar a esa misma clase o grupo.

No obstante, puede ocurrir que una regla de pertenencia o asociación de clases sea muy precisa (métrica PRCR (R) con valor alto), pero tener una baja cobertura de casos a los que se la puede aplicar en el modelo construido (métrica COBER (R) con valor bajo). En consecuencia, termina resultando una regla irrelevante y de poco interés para el proyecto de explotación de información.

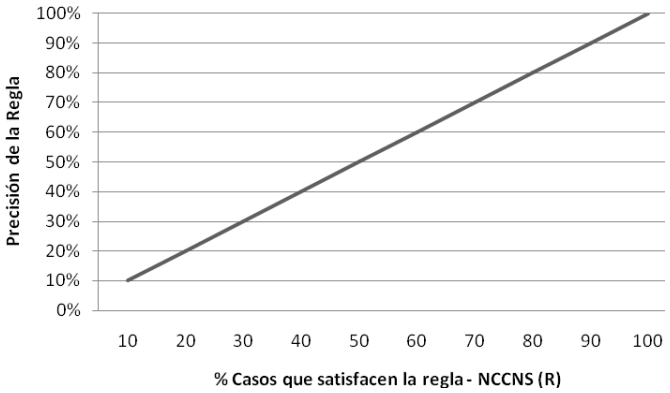

Fig. 18. Proporción de casos cubiertos por una regla para la métrica PRCR (R)

Del análisis realizado se concluye que, toda regla R_i cuya métrica de precisión PRCR (R_i) supera el 60-70% de certeza, se considera confiable y de interés para asociar nuevos casos a esa clase o grupo, facilitando el descubrimiento de conocimiento para el proyecto de explotación de información.

d) Regla Experimental

Como regla experimental del comportamiento de la métrica PRCR (R) analizada, se concluye que la misma aumenta su valor cuando se incrementa el número de casos a los que se puede aplicar una regla de pertenencia a una clase o grupo, asociada a la métrica NCCNS (R), respecto del total de casos que sólo satisfacen la precondición o antecedente de esa misma regla, independientemente de la clase que asocia. Mientras que, cuando el número de casos que a los que se le puede aplicar la regla disminuye, también lo hace la métrica de precisión de la regla.

Por otra parte, si la métrica de precisión de una regla supera un valor mínimo de umbral, esa regla resultaría confiable para utilizarla en el proyecto de explotación de información al permitir obtener conocimiento de su precondición o antecedente.

El valor de umbral propuesto para la métrica puede ser refinado si se desea obtener una mayor certeza en la regla de pertenencia o asociación a una clase o grupo del modelo. No obstante, para que una regla sea relevante y de interés para un proyecto de explotación de información, no sólo interesa que la misma tenga mucha precisión, sino también que cubra un amplio número de casos, asociado a la métrica COBER (R).

8. Métrica de Usabilidad de Atributos del Modelo

Para estudiar la métrica de Usabilidad de Atributos de un modelo asociado al Descubrimiento de Reglas de pertenencia de clases o grupos – USAT (M), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de la misma.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en la métrica USAT (M), considerada como métrica derivada. La lista de variables experimentales se muestra en la Tabla CII.

TABLA CII. VARIABLES EXPERIMENTALES PARA LA MÉTRICA USAT (M)

Variable Experimental	Descripción
NA (M)	Número de atributos utilizados para construir el modelo.
NAPR (M)	Número de atributos distintos sobre los cuales las reglas de pertenencia a cada clase o grupo del modelo imponen condiciones.
USAT (M)	Usabilidad de los atributos del modelo, la cual se define como: $\text{USAT(M)} = \frac{\text{NAPR(M)}}{\text{NA(M)}}$

b) Diseño Experimental

Para analizar el comportamiento de la métrica enunciada, se utiliza un banco de pruebas simulado de 80 modelos de explotación de información con la información obtenida de las reglas de pertenencia a clases generadas, luego de aplicar un algoritmo de inducción TDIDT [68] al modelo. Para ello, se define un rango de valores específico para cada una de las variables experimentales independientes, considerando las restricciones indicadas en [65] según el tamaño del proyecto, generando diferentes combinaciones sujetas a análisis. Los valores simulados para las variables experimentales independientes se muestran en la Tabla CIII.

TABLA CIII. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LA MÉTRICA USAT (M)

Variable Experimental	Descripción
NA (M)	Número de casos a utilizar para una clase o grupo del modelo durante su entrenamiento, con un rango de valores específico 5, 10, 20, 30, 40, 50, 60 y 70.
NAPR (M)	Número de casos que satisfacen la aplicación de la regla, con un rango de valores específico 10, 20, 30, 40, 50, 60, 70, 80, 90 y 100. Este rango de valores se expresa como un porcentaje del número total de casos a utilizar en la clase o grupo.

El resultado y comportamiento de la métrica derivada USAT (M) se obtiene luego de aplicar un modelo de Descubrimiento de Reglas al modelo construido. Este comportamiento está asociado al número total de atributos distintos sobre los cuales las reglas de pertenencia a cada clase o grupo imponen condiciones, respecto del total de atributos utilizados para construir el modelo de explotación de información, y cuyas métricas básicas NAPR (M) y NA (M), se indican en la Tabla CIII.

En el caso de la métrica NAPR (M), al estar sometida a un proceso de simulación, no se le puede asignar un número a priori de atributos a utilizar. Por tal motivo, se decide establecer un rango de valores porcentuales con intervalos de a 10%, definido entre 10 y 100% que represente el número de atributos diferentes, obtenidos en las reglas, luego de aplicar un modelo de descubrimiento de reglas al proyecto, respecto del número total de atributos utilizados en la construcción del modelo. A partir de los rangos de valores definidos, se analiza el comportamiento general de la métrica USAT (M).

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de la métrica USAT (M), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que el valor de la métrica USAT (M) aumenta cuando el porcentaje de atributos sobre el que las reglas descubiertas imponen condiciones de pertenencia a clases o grupos, asociada a la métrica NAPR (M), toma valores entre 10% y 100% respecto del total de atributos utilizados para construir el modelo de explotación de información, asociado a la métrica NA (M).

Por el contrario, cuando el porcentaje de atributos correspondiente a la métrica NAPR (M) disminuye, manteniendo la misma cantidad de atributos en el modelo construido, el valor de la métrica USAT (M) también disminuye. En la Fig. 19 se muestra la proporción de atributos utilizados en un modelo de descubrimiento de reglas en función del número de atributos sobre el que las reglas descubiertas imponen condiciones de pertenencia a clases o grupos.

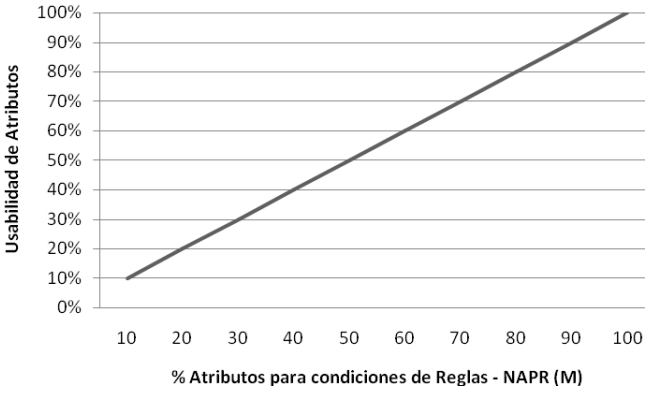

Fig. 19. Proporción de atributos usados en un modelo de descubrimiento de reglas para la métrica USAT (M)

d) Regla Experimental

Como regla experimental del comportamiento de la métrica USAT (M) analizada, se concluye que la misma aumenta su valor cuando se incrementa el número de atributos sobre el que las reglas descubiertas establecen condiciones de pertenencia a clases o grupos, asociada a la métrica NAPR (M), respecto del total de atributos utilizados para construir el modelo de explotación de información, asociado a la métrica NA (M). Caso contrario, el valor de la métrica de usabilidad de atributos disminuye.

Para un proyecto de explotación de información, un valor ALTO de la métrica USAT (M) implica que todos los atributos utilizados en la construcción del modelo son relevantes y necesarios para encontrar las reglas que describen el

comportamiento de las clases o grupos. Mientras que un valor BAJO de la métrica, es un indicador de posibles problemas en el desarrollo del modelo de explotación de información, ya que refleja que el mismo se construye con atributos que no son todos relevantes y necesarios.

9. Métrica de Grado de Incidencia de Atributos

Para estudiar la métrica de Grado de Incidencia de los Atributos sobre un atributo clase en un modelo asociado al Descubrimiento de Dependencias Significativas – GINC (A_{vi}), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de la misma.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en la métrica GINC (A_{vi}), considerada como métrica derivada. La lista de variables experimentales se muestra en la Tabla CIV.

TABLA CIV. VARIABLES EXPERIMENTALES PARA LA MÉTRICA GINC (A_{vi})

Variable Experimental	Descripción
NCAT (A_{Vi})	Número de casos cubiertos por una clase o grupo del atributo clase, cuando un atributo significativo A toma el valor V_i.
NCE (C)	Número de casos a utilizar por cada clase para en el entrenamiento del modelo. Para un problema de N clases, representa la cantidad de casos de entrenamiento utilizados por cada clase C_i del modelo.
GINC (A_{Vi})	Grado de incidencia que cada valor V_i de un atributo significativo A tiene sobre una clase o grupo del atributo clase, la cual se define como: $GINC(A_{Vi}) = \frac{NCAT(A_{Vi})}{NCE(C)}$

b) Diseño Experimental

Para analizar el comportamiento de la métrica enunciada, se utiliza un banco de pruebas simulado de 80 modelos de explotación de información con la información obtenida de las reglas de pertenencia a clases generadas, luego de aplicar un algoritmo de inducción TDIDT [68] al modelo. Para ello, se define un rango de valores específico para cada una de las variables experimentales independientes, considerando las restricciones indicadas en [65] según el tamaño del proyecto, generando diferentes combinaciones sujetas a análisis. Los valores simulados para las variables experimentales independientes se muestran en la Tabla CV.

TABLA CV. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LA MÉTRICA GINC (A_{vi})

Variable Experimental	Descripción
NCAT (A_{Vi})	Número de casos cubiertos por una clase o grupo del atributo clase, cuando un atributo significativo A toma el valor V_i, con un rango de valores específico 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 y 100. Este rango de valores se expresa como un porcentaje del número total de casos cubiertos por la clase o grupo.
NCE (C)	Número de casos a utilizar para una clase o grupo del modelo durante su entrenamiento, con un rango de valores específico 1.000, 2.000, 3.000 y 4.000.

El resultado y comportamiento de la métrica derivada GINC (A_{vi}) se obtiene luego de aplicar un modelo de Descubrimiento de Dependencias Significativas al modelo construido. Este comportamiento está asociado al número de casos de entrenamiento que son cubiertos para un valor (clase o grupo) del atributo clase del modelo, cuando un atributo significativo A toma el valor V_i, respecto del total de casos que pertenecen a esa misma clase, y cuyas métricas básicas NCAT (A_{vi}) y NCE (C), se indican en la Tabla CV.

Por otra parte, se tiene como restricción que la cantidad de casos por cada valor asociado al atributo significativo debe cubrir la totalidad de los casos de entrenamiento que contiene el atributo clase del modelo.

$$NCE(C) = \sum_{i=1}^{n} A(v_i)$$

En el caso de la métrica NCAT (A_{vi}), al estar sometida a un proceso de simulación, no se le puede asignar un número a priori de casos a utilizar. Además, debe estar asociada al número de valores diferentes que puede tomar el atributo significativo al que se le aplica la métrica. Por tal motivo, se decide establecer un rango de valores porcentuales con intervalos de a 10%, definido entre 0 y 100% que represente el número de casos seleccionados de un valor (clase o grupo) del atributo clase, cuando el atributo significativo toma cada valor V_i, respecto del número total de casos que contiene la clase del modelo. Asimismo, se considera para este análisis que el atributo significativo tiene asociados 2 y 3 valores, respectivamente. A partir de los rangos de valores definidos, se analiza el comportamiento general de la métrica GINC (A_{vi}).

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de la métrica GINC (A_{vi}), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que para un atributo significativo con dos valores *v1* y *v2* asociados, el valor de la métrica GINC (A_{vi}) aumenta cuando el porcentaje de casos cubiertos por alguno de los dos valores, para un valor (clase o grupo) del atributo clase, tiende al 100% del total de casos de entrenamiento de la clase analizada (para el otro valor del atributo significativo la métrica disminuye). En la Fig. 20 se muestra la variación del grado de incidencia que cada valor *v1* y *v2* ejerce sobre el valor (clase o grupo) del atributo clase, en función de la proporción de casos cubiertos por el mismo cuando el atributo significativo toma cada uno de estos valores.

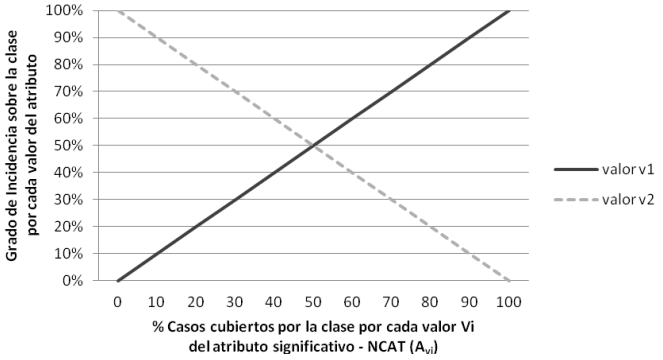

Fig. 20. Incidencia de un atributo A con dos valores según la proporción de casos cubiertos por la clase para la métrica GINC (A_{vi})

En un segundo análisis experimental, se observa que para un atributo significativo con tres valores v1, v2 y v3 asociados, el valor de la métrica GINC (Avi) aumenta cuando el porcentaje de casos cubiertos por alguno de los tres valores, para un valor (clase o grupo) del atributo clase, tiende al 100% del total de casos de entrenamiento de la clase analizada (para

los otros dos valores del atributo significativo la métrica disminuye).

En las Fig. 21, 22 y 23, se muestran las variaciones del grado de incidencia que cada valor v1, v2 y v3 ejerce sobre el valor (clase o grupo) del atributo clase, en función de la proporción de casos cubiertos por el mismo cuando la métrica NCAT (v1) del atributo significativo toma valores 10%, 40% y 70%, respectivamente.

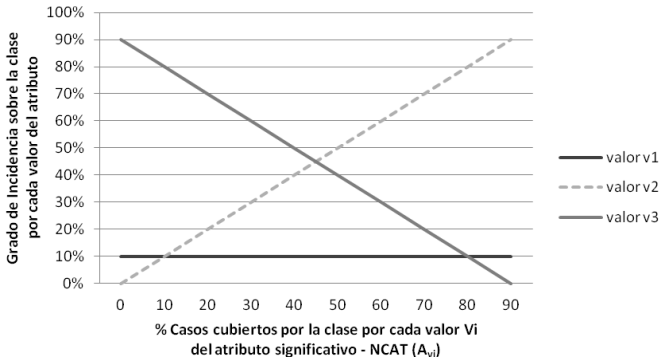

Fig. 21. Incidencia de un atributo A con tres valores según la proporción de casos cubiertos por la clase cuando NCAT (v1)=10%, para la métrica GINC (A_{vi})

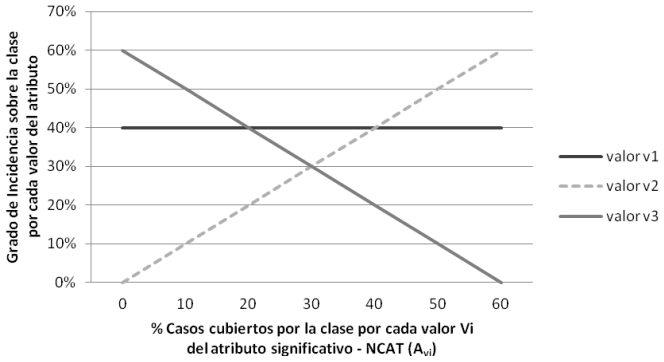

Fig. 22. Incidencia de un atributo A con tres valores según la proporción de casos cubiertos por la clase cuando NCAT (v1)=40%, para la métrica GINC (A_{vi})

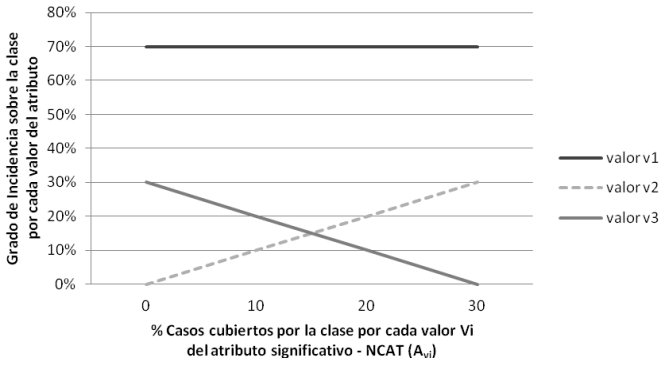

Fig. 23. Incidencia de un atributo A con tres valores según la proporción de casos cubiertos por la clase cuando NCAT (v1)=70%, para la métrica GINC (A_{vi})

Del análisis realizado a la métrica GINC (Avi), se concluye que cuando aumenta el número de casos cubiertos por alguno de los valores asociados a un atributo significativo, para una clase o grupo del atributo clase, el valor de la métrica analizada converge a 1 (uno) para ese valor de atributo significativo, mientras que para los demás valores la métrica converge a 0 (cero).

Cabe resaltar, que un modelo de Descubrimiento de Dependencias Significativas asume que las relaciones de dependencias entre los atributos significativos son condicionalmente independientes entre sí dado un atributo clase [20] [42]. Por consiguiente, las conclusiones obtenidas de este análisis son aplicables a cualquier número de atributos significativos que tenga el modelo de explotación de información construido.

d) Regla Experimental

Como regla experimental del comportamiento de la métrica analizada, se concluye que todo valor asociado a un atributo significativo del modelo construido, y cuya métrica GINC (A_{vi}) sea máxima para ese valor, representa las características o factores que mayor incidencia tienen sobre un determinado resultado de un problema (representado por el atributo clase) en el proyecto de explotación de información.

Cabe mencionar que, cuando la métrica GINC (Avi) toma valores muy similares en dos o más valores del atributo significativo, implica que no hay una única característica dentro de este atributo que tenga más influencia sobre el atributo clase.

E. Estudio de las Métricas de Proyectos

Para estudiar el comportamiento de las Métricas de Proyectos, se decide generar un banco de pruebas simulado con diferentes cantidades de proyectos de explotación de información, considerando las restricciones indicadas en [64] para un proyecto de tamaño pequeño, al que se le aplican las métricas propuestas para la Evaluación del Proceso de Desarrollo (apartado IV, sección *B*, sub-sección *3.a*) y Entrega del Proyecto (apartado IV, sección *B*, sub-sección *3.b*).

La simulación de las Métricas de Proyectos utiliza las siguientes variables independientes (sub-sección *1*) y dependientes (sub-sección *2*).

1. Variables Independientes

Las variables independientes que se van a generar mediante el proceso de simulación, son las correspondientes a las métricas básicas definidas en la secciones de Evaluación del Proceso de Desarrollo (apartado IV, sección *B*, sub-sección *3.a*) y Entrega del Proyecto (apartado IV, sección *B*, sub-sección *3.b*) y que afectan directamente a las métricas derivadas. Para estas métricas básicas se define un valor específico o un valor aleatorio, considerando las restricciones por el tamaño del proyecto, restringiendo así la cantidad de combinaciones. Las variables independientes a ser utilizadas se muestran en la Tabla CVI.

2. Variables Dependientes

Para este proceso de simulación las variables dependientes, o sea las que son afectadas por las variables independientes, son los resultados de aplicar las fórmulas de las métricas derivadas de Éxito de Resultados del Proceso de Desarrollo (sub-sección *3*) y Desvío en el Esfuerzo del Proyecto (sub-sección *4*) para las variables independientes definidas. Las variables dependientes a ser utilizadas se muestran en la Tabla CVII.

Cabe mencionar, que algunas de las métricas de proyectos sólo se calculan en función de la suma de los valores de otras métricas básicas y no se requiere estudiar su comportamiento. Estas métricas denominadas por la norma ISO/IEC 9126 como de agregación se indican en la Tabla CVIII.

TABLA CVI. VARIABLES INDEPENDIENTES PARA MÉTRICAS DE PROYECTOS

Variable Independiente (métrica básica)	Descripción
EFZE (P)	Esfuerzo estimado para desarrollar el proyecto de explotación de información, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs. Cabe mencionar que si bien esta no es una métrica básica, sino el esfuerzo estimado al inicio del proyecto, se la utiliza como variable independiente para las Métricas de Proyectos.
EFZR (P)	Esfuerzo real aplicado para desarrollar el proyecto de explotación de información, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NMIA	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Alto, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NMIB	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Bajo, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NMIM	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Medio, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs.
NMOD	Número de modelos construidos para el proyecto de explotación de información, con un valor específico o un valor aleatorio del rango de valores predefinido para un proyecto de tamaño pequeño y de aplicación en PyMEs. Cabe mencionar que si bien esta métrica básica se define como una métrica de Modelado, se la utiliza como variable independiente para las Métricas de Proyectos.

TABLA CVII. VARIABLES DEPENDIENTES PARA MÉTRICAS DE PROYECTOS

Variable Dependiente (métrica derivada)	Descripción
EPD (P)	Éxito de resultados del proceso de desarrollo para el proyecto de explotación de información, el cual depende de la cantidad de modelos construidos probados, y cuyo nivel de confiabilidad de resultados obtenidos sea Bajo, Medio y Alto.
DEFZ (P)	Desvío en el esfuerzo para desarrollar el proyecto de explotación de información, el cual depende del error relativo entre el esfuerzo estimado al inicio del proyecto y el esfuerzo real aplicado para desarrollar las tareas en el proyecto.

La relación entre las variables independientes y dependientes indicando como afectan unas a otras puede verse en la Fig. 24.

3. Métrica de Éxito de Resultados del Proceso de Desarrollo

Para estudiar la métrica de Éxito de Resultados obtenidos en el proceso de desarrollo del proyecto de explotación de información – EPD (P), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de la misma.

TABLA CVIII. VARIABLES DEPENDIENTES PARA MÉTRICAS DE PROYECTOS NO CONSIDERADAS EN EL ANÁLISIS

Variable Dependiente (métrica agregación)	Descripción
DRPY (P)	Duración real del proyecto de explotación de información, el cual se define como: $DRPY(P) = \sum TRSubprc_i$ Donde $TRSubprc_i$ es el tiempo real insumido para desarrollar las tareas de cada uno de los subprocesos Entendimiento del Negocio, Entendimiento de los Datos, Preparación de los Datos, Modelado, Evaluación y Entrega [84]. Este tiempo se relaciona con el progreso del proyecto.
PRGS (P)	Progreso del proyecto de explotación de información, el cual se define como: $PRGS(P) = TESubprc_i - TRSubprc_i$ donde $TESubprc_i$ y $TRSubprc_i$ es el tiempo estimado y real insumido para desarrollar las tareas de cada uno de los subprocesos Entendimiento del Negocio, Entendimiento de los Datos, Preparación de los Datos, Modelado, Evaluación y Entrega [85]. ▪ Si PRGS (P) > 0 indica que el proyecto está avanzando dentro de los tiempos planificados y estimados. No obstante, se debe controlar que no se produzca una sobreestimación de tiempos en el desarrollo de las tareas. ▪ Si PRGS (P) < 0 indica que el proyecto está atrasado y que se subestimó el tiempo estimado para desarrollar las tareas del proyecto. La sobreestimación o subestimación de tiempos impacta en el esfuerzo real requerido y en la métrica del desvío en el esfuerzo DEFZ (P) para realizar el proyecto.

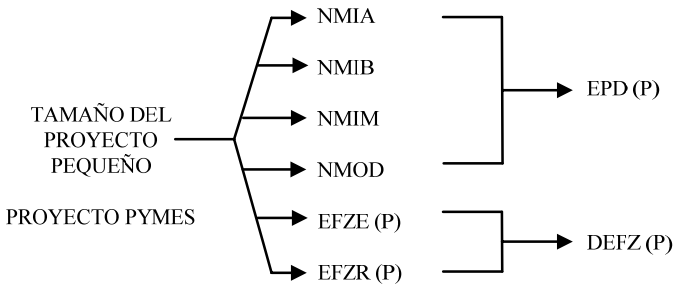

Fig. 24. Relación entre las variables independientes y dependientes para Métricas de Proyectos

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en la métrica EPD (P), considerada como métricas derivada. La lista de variables experimentales se muestra en la Tabla CIX.

b) Diseño Experimental

Para analizar el comportamiento de la métrica enunciada, se utiliza un banco de pruebas simulado de 55 proyectos de explotación de información, definiendo un rango de valores específico para cada una de las variables experimentales independientes y considerando las restricciones indicadas en [65] según el tamaño del proyecto. Además, se considera que se aplica a modelos de explotación de información entrenados y probados, con un alto nivel de exactitud en la clasificación de clases y un nivel alto y/o medio de precisión.

Estas consideraciones corresponden a los criterios que normalmente se toman en cuenta al momento de evaluar la

calidad de los modelos construidos para un proyecto de explotación de información real. De esta manera, se generan diferentes combinaciones sujetas a análisis.

Los valores simulados para las variables experimentales independientes se muestran en la Tabla CX.

TABLA CIX. VARIABLES EXPERIMENTALES PARA LA MÉTRICA EPD (P)

Variable Experimental	Descripción
EPD (P)	Éxito de resultados del proceso de desarrollo para el proyecto de explotación de información, el cual se define como: $EPD(P) = \frac{\sum NMI_i * Peso(i)}{NMOD}$ donde NMI_i es el número de modelos de explotación de información cuyo interés *i* de resultados es Alto, Medio o Bajo
NMIA	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Alto.
NMIB	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Bajo.
NMIM	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Medio.
NMOD	Número de modelos construidos para el proyecto de explotación de información.

TABLA CX. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LA MÉTRICA EPD (P)

Variable Experimental	Descripción
NMIA	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Alto, con un rango de valores específico 0, 1, 2, 3, 4 y 5.
NMIB	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Bajo, con un rango de valores específico 0, 1, 2, 3, 4 y 5.
NMIM	Número de modelos de explotación de información cuyo interés y utilidad de los resultados es Medio, con un rango de valores específico 0, 1, 2, 3, 4 y 5.
NMOD	Número de modelos construidos para el proyecto de explotación de información, con un rango de valores específico 1, 2, 3, 4 y 5, según lo predefinido para un proyecto de tamaño pequeño.

El resultado y comportamiento de la métrica derivada EPD (P) está asociado al número de modelos probados y analizados con el usuario para el proyecto de explotación de información, cuyos resultados obtenidos, en relación al conocimiento descubierto y utilidad, sean de interés Bajo, Medio o Alto para el usuario del proyecto, respecto del total de los modelos construidos. Las métricas básicas NMA, NMB, NMM, NMOD asociadas a estos resultados, se indican en la Tabla CX.

Por otra parte, en [65] se menciona que en un proyecto de explotación de información para PyMEs, la cantidad de modelos construidos no suele ser superior a 6 modelos. Por consiguiente, se tiene como restricción que la cantidad de modelos probados y analizados con el usuario, independientemente del nivel de interés de los resultados obtenidos, debe cubrir la totalidad de los modelos construidos.

$$NMOD = NMIB + NMIM + NMIA$$

A partir de los rangos de valores definidos, se generan los datos de los modelos de explotación de información probados y analizados con el usuario. De esta manera, se busca obtener el comportamiento general de la métrica de éxito de resultados del proceso de desarrollo del proyecto EPD (P) (variable experimental dependiente), en función de la variación de estas variables.

En virtud que no existe un criterio universal que defina qué indicador se debe adoptar para considerar un proceso de desarrollo con nivel ALTO, MEDIO o BAJO de éxito en los resultados obtenidos del análisis a los modelos, en la Tabla CXI se indica los valores y rangos contemplados para un proyecto de explotación de información para PyMEs, considerados a partir de la propia experiencia.

TABLA CXI. RANGOS Y VALORES RELACIONADOS A LA MÉTRICA EPD (P)

Rango	Descripción
Bajo	Menos de 60% de éxito de resultados en el proceso de desarrollo
Medio	Entre 61 y 80% de éxito de resultados en el proceso de desarrollo
Alto	Más de 80% de éxito de resultados en el proceso de desarrollo

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de la métrica EPD (P), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que la métrica derivada EPD (P) toma un valor ALTO, considerándose como parámetro para este análisis un valor relativo superior al 80%, cuando el número de modelos de un proyecto de explotación de información con nivel Alto de interés en los resultados obtenidos, asociado a la métrica NMIA, está entre 1 y 5. A su vez, el número de modelos con nivel de interés Medio, asociado a la métrica NMIM, está entre 0 y 2, mientras que los de interés Bajo, asociado a la métrica NMIB, están entre 0 y 1. En las Fig. 25, 26 y 27, se muestra la proporción de proyectos simulados cuya métrica de éxito del proceso de desarrollo EPD (P) es ALTA, en función del número de modelos necesarios por nivel de interés de resultados.

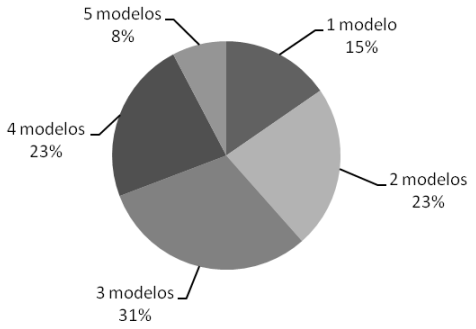

Fig. 25. Distribución de proyectos con alto interés de resultados del modelo para la métrica EPD (P) = Alto

En la Fig. 25, se observa que si bien todo proyecto debe tener al menos un modelo con alto interés de resultados para el usuario, es a partir del segundo modelo, asociado a la métrica NMIA, que la métrica de éxito de resultados EPD (P) es más representativa con un valor ALTO.

Cabe aclarar que, si un proyecto sólo tiene un modelo con alto interés de resultados, hay una probabilidad del 87% que la métrica EPD (P) tome un valor MEDIO o BAJO. Por tal motivo, se excluyen los proyectos con un solo modelo de alto interés como condición para que la métrica de éxito de resultados EPD (P) tome un valor ALTO. A su vez, si en un proyecto de explotación de información todos sus modelos son de alto interés para el usuario, se obtiene el valor máximo para la métrica de éxito del proceso de desarrollo EPD (P), pero sólo el 8% de los proyectos simulados con ALTO valor de la métrica EPD (P) representan esta condición.

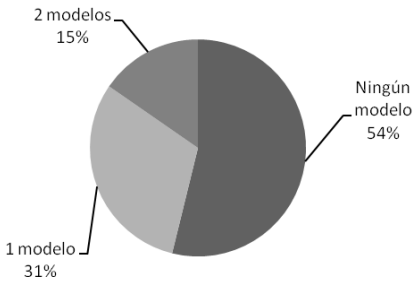

Fig. 26. Distribución de proyectos con interés medio de resultados del modelo para la métrica EPD (P) = Alto

En la Fig. 26, y considerando la conclusión obtenida con la métrica NMIA, se observa que para que la métrica de éxito de resultados EPD (P) tome un valor ALTO, un proyecto no debe tener modelos que representen para el usuario un interés medio en los resultados obtenidos, asociado a la métrica NMIM. Esto representa el 40% de los proyectos simulados. Sin embargo, si la métrica NMIN toma valores 1 o 2, el porcentaje de proyectos con valor ALTO de la métrica EPD (P) es aproximadamente el mismo. De este análisis se concluye que la métrica NMIN puede ser menor o igual a 2 modelos, para que el éxito de resultados sea ALTO.

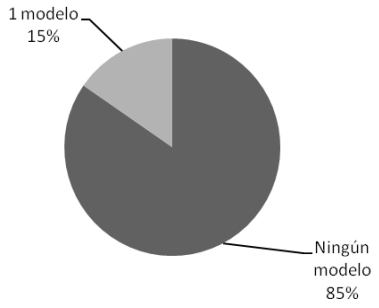

Fig. 27. Distribución de proyectos con interés bajo de resultados del modelo para la métrica EPD (P) = Alto

En la Fig. 27, y considerando las conclusiones obtenidas con las métricas NMIA y NMIM, se observa que para que la métrica de éxito de resultados EPD (P) tome un valor ALTO, un proyecto no debe tener modelos que representen, para el usuario experto, un bajo interés en los resultados obtenidos, asociado a la métrica NMIB.

Del análisis realizado a los modelos de un proyecto de explotación de información, cuando la métrica EPD (P) toma un valor ALTO, se concluye que cuando el número de modelos con Alto nivel de interés en los resultados obtenidos, asociado a la métrica NMIA, es mayor o igual a 2, el número de modelos con nivel de interés Medio, asociado a la métrica NMIM, es menor o igual a 2, y el número de modelos con nivel de interés Bajo, asociado a la métrica NMIB, es 0, se obtiene la mayor cantidad de proyectos con características similares y un ALTO nivel de éxito en los resultados del proceso de desarrollo. Esto representa el 70% de los proyectos simulados, obteniéndose un valor medio relativo para la métrica EPD (P) del 92.5%.

En un segundo análisis experimental, se observa que la métrica derivada EPD (P) toma un valor MEDIO, considerándose como parámetro para este análisis un valor relativo entre 61 y 80%, cuando el número de modelos de un proyecto de explotación de información con nivel Alto de interés en los resultados obtenidos, asociado a la métrica NMIA, está entre 1 y 3. A su vez, el número de modelos con nivel de interés Medio, asociado a la métrica NMIM, está entre 0 y 4, mientras que los de interés Bajo, asociado a la métrica NMIB, están entre 0 y 2. En las Fig. 28, 29 y 30, se muestra la proporción de proyectos simulados cuya métrica de éxito del proceso de desarrollo EPD (P) es de valor MEDIO, en función del número de modelos necesarios por nivel de interés de resultados.

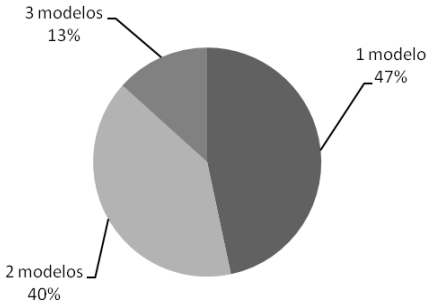

Fig. 28. Distribución de proyectos con interés alto de resultados del modelo para la métrica EPD (P) = Medio

En la Fig. 28, se observa que para que la métrica de éxito de resultados EPD (P) tome un valor MEDIO, un proyecto debe tener 1 o 2 modelos que representen para el usuario un alto interés en los resultados obtenidos, asociado a la métrica NMIA. Esto representa el 87% de los proyectos simulados, obteniéndose un valor medio relativo para la métrica EPD (P) del 68%.

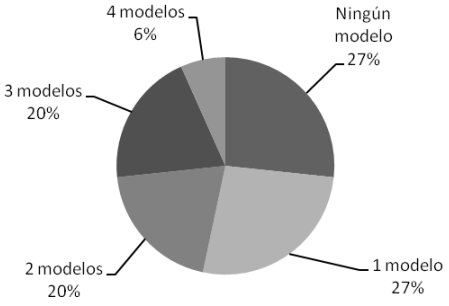

Fig. 29. Distribución de proyectos con interés medio de resultados del modelo para la métrica EPD (P) = Medio

En la Fig. 29, y considerando la conclusión obtenida con la métrica NMIA, se observa que para que la métrica de éxito de resultados EPD (P) tome un valor MEDIO, un proyecto debe tener hasta 3 modelos que representen para el usuario un interés medio en los resultados obtenidos, asociado a la métrica NMIM. Esto representa el 80% de los proyectos simulados con un valor medio relativo para la métrica EPD (P) del 68%. No obstante, si se considera un proyecto con 4 modelos de interés medio de resultados, se cubre el 86% de proyectos con un valor medio para la métrica EPD (P) del 68%. Con lo cual, tomando en cuenta un proyecto con esta cantidad de modelos e interés medio de resultados, asociado a la métrica NMIM, los resultados no varían significativamente.

En la Fig. 30, y considerando las conclusiones obtenidas con las métricas NMIA y NMIM, se observa que para que la métrica de éxito de resultados EPD (P) tome un valor MEDIO, un proyecto debe tener un modelo que represente para el usuario un bajo interés en los resultados obtenidos, asociado a la métrica NMIB. No obstante, cabe mencionar que, si un proyecto considera que no hay modelos con bajo interés de

resultados, existe un 55% de probabilidades que la métrica EPD (P) tome un valor ALTO, mientras que si se consideran dos modelos, existe un 60% de probabilidades que la métrica EPD (P) tome un valor BAJO.

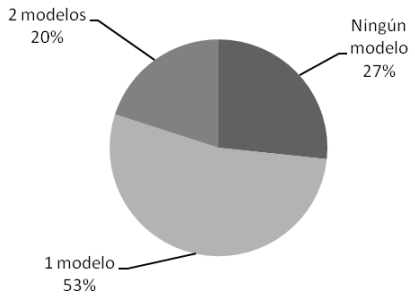

Fig. 30. Distribución de proyectos con bajo interés de resultados del modelo para la métrica EPD (P) = Medio

Analizando ambas situaciones, si se toma un proyecto con hasta un modelo de bajo interés de resultados, se cubre el 73% de los proyectos simulados, con un valor medio relativo para la métrica EPD (P) del 69%. Mientras que tomando 1 ó 2 modelos, se cubre el 60% de los proyectos simulados, con un valor medio para la métrica EPD (P) del 68%. Por tal motivo, se prefiere excluir los proyectos con dos modelos de bajo interés como condición para que la métrica de éxito de resultados EPD (P) tome un valor MEDIO.

Del análisis realizado a los modelos de un proyecto de explotación de información, cuando la métrica EPD (P) toma un valor MEDIO, se concluye que cuando el número de modelos con Alto nivel de interés en los resultados obtenidos, asociado a la métrica NMIA, es 1 o 2, el número de modelos con nivel de interés Medio, asociado a la métrica NMIM, es menor o igual a 4, y los de interés Bajo, asociado a la métrica NMIB, es a lo sumo 1, se obtiene la mayor cantidad de proyectos con características similares y un nivel MEDIO de éxito en los resultados del proceso de desarrollo. Esto representa el 73% de los proyectos simulados, obteniéndose un valor medio relativo para la métrica EPD (P) del 69%.

En el último análisis experimental, se observa que la métrica derivada EPD (P) toma un valor BAJO, considerándose como parámetro para este análisis un valor relativo hasta el 60%, cuando el número de modelos de un proyecto de explotación de información con nivel Alto de interés en los resultados obtenidos, asociado a la métrica NMIA, está entre 0 y 2. A su vez, el número de modelos con nivel de interés Medio y Bajo, asociados a las métricas NMIM y NMIB, está entre 0 y 5, respectivamente. En las Fig. 31, 32 y 33, se muestra la proporción de proyectos simulados cuya métrica de éxito del proceso de desarrollo EPD (P) es de valor BAJO, en función del número de modelos necesarios por nivel de interés de resultados.

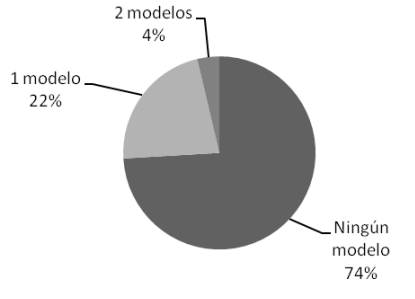

Fig. 31. Distribución de proyectos con alto interés de resultados del modelo para la métrica EPD (P) = Bajo

En la Fig. 31, se observa que para que la métrica de éxito de resultados EPD (P) tome un valor BAJO, un proyecto no debe tener modelos que representen para el usuario un alto interés en los resultados obtenidos, asociado a la métrica NMIA. Sin embargo, si se considera que un proyecto también puede tener un modelo con bajo interés de resultados, existe un 40% de probabilidades que la métrica EPD (P) tome un valor BAJO. No obstante, tomando un proyecto con hasta un modelo de alto interés, se cubre el 96% de los proyectos simulados.

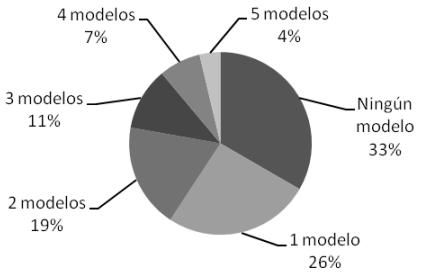

Fig. 32. Distribución de proyectos con interés medio de resultados del modelo para la métrica EPD (P) = Bajo

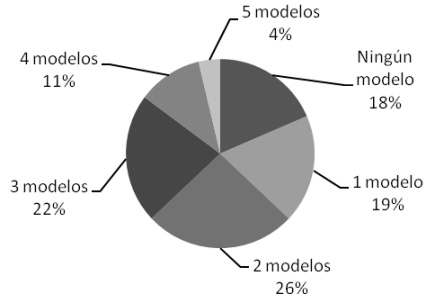

Fig. 33. Distribución de proyectos con interés bajo de resultados del modelo para la métrica EPD (P) = Bajo

En la Fig. 32, y considerando la conclusión obtenida con la métrica NMIA, se observa que para que la métrica de éxito de resultados EPD (P) tome un valor BAJO, un proyecto puede tener cualquier cantidad de modelos que representen para el usuario un interés medio en los resultados obtenidos, asociado a la métrica NMIM. Esto se debe a que no hay un sesgo marcado dentro de los proyectos simulados, que oriente hacia un número específico de modelos a considerar. De este análisis se concluye que la métrica NMIN no es representativa para que el éxito de resultados sea BAJO, ya que depende de la cantidad de modelos con alto y bajo interés de resultados en el proyecto.

En la Fig. 33, y considerando las conclusiones obtenidas con las métricas NMIA y NMIM, se observa que para que la métrica de éxito de resultados EPD (P) tome un valor BAJO, un proyecto también puede tener cualquier cantidad de modelos que representen para el usuario un bajo interés en los resultados obtenidos, asociado a la métrica NMIB. Sin embargo, si un proyecto considera que puede tener hasta un modelo con bajo interés de resultados, existe un 50% de probabilidades que la métrica EPD (P) tome un valor BAJO. Por tal motivo, se prefiere excluir estas cantidades de modelos como condición para que la métrica de éxito de resultados EPD (P) tome un valor BAJO.

Del análisis realizado a los modelos de un proyecto de explotación de información, cuando la métrica EPD (P) toma un valor BAJO, se concluye que cuando el número de modelos con Alto nivel de interés en los resultados obtenidos, asociado a la métrica NMIA, es a lo sumo 1 y el número de modelos con Bajo nivel de interés, asociado a la métrica NMIB, es mayor o

igual a 2, se obtiene la mayor cantidad de proyectos con características similares y un nivel BAJO de éxito en los resultados del proceso de desarrollo. Esto representa el 60% de los proyectos simulados, obteniéndose un valor medio relativo para la métrica EPD (P) del 42%.

d) Regla Experimental

Como regla experimental del comportamiento de la métrica analizada, se concluye que:

- El éxito de resultados del proceso de desarrollo de un proyecto de explotación de información es ALTO, cuando el número de modelos que presentan un alto interés para el usuario, por el conocimiento descubierto y su utilidad, asociado a la métrica NMIA, es mayor o igual a 2. A su vez, el número de modelos que presentan un interés medio, asociado a la métrica NMIM, es menor o igual a 2, mientras que el número de modelos con bajo interés para el usuario, asociado a la métrica NMIB, es a lo sumo 1. Esto representa el 70% de los proyectos simulados y un valor medio para la métrica EPD (P) del 92,5%, lo que indica que el proceso seguido para el desarrollo del proyecto cumple satisfactoriamente con los criterios de éxito definidos para el mismo. Cuanto más alto sea el número de modelos con alto interés de resultados obtenidos, más se acerca la métrica de éxito EPD (P) a su valor máximo (100%). En la Fig. 34 se muestra el número estimado de modelos necesarios y su interés de resultados para un proyecto, por cada métrica analizada.
- El éxito de resultados del proceso de desarrollo de un proyecto de explotación de información es MEDIO, cuando el número de modelos que presentan un alto interés para el usuario, por el conocimiento descubierto y su utilidad, asociado a la métrica NMIA, es 1 o 2. A su vez, el número de modelos que presentan un interés medio, asociado a la métrica NMIM, es menor o igual a 4, mientras que el número de modelos con bajo interés para el usuario, asociado a la métrica NMIB, es a lo sumo 1. Esto representa el 73% de los proyectos simulados y un valor medio para la métrica EPD (P) del 69%, lo que indica que se cumplen parcialmente con los criterios de éxito definidos para el proyecto. En consecuencia, se necesita revisar y ajustar alguna/s de la/s tarea/s del proceso de desarrollo. Cabe mencionar que, cuando la métrica NMIM toma valor 5, el éxito del proceso pasa a ser BAJO. En la Fig. 34 se muestra el número estimado de modelos necesarios y su interés de resultados para un proyecto, por cada métrica analizada.
- El éxito de resultados del proceso de desarrollo de un proyecto de explotación de información es BAJO, cuando el número de modelos que presentan un alto interés para el usuario, por el conocimiento descubierto y su utilidad, asociado a la métrica NMIA, es a lo sumo 1. Mientras que el número de modelos con bajo interés para el usuario, asociado a la métrica NMIB, es mayor o igual a 2. Esto representa el 60% de los proyectos simulados y un valor medio para la métrica EPD (P) del 42%, lo que indica que no se cumplen con los criterios de éxito definidos para el proyecto. En consecuencia, se debe analizar y replantear las tareas del proceso de desarrollo inherentes al entendimiento y preparación de los datos y modelado del proyecto, según su tarea de descubrimiento. En la Fig. 34 se muestra el número estimado de modelos necesarios y su interés de resultados para un proyecto, por cada métrica analizada.

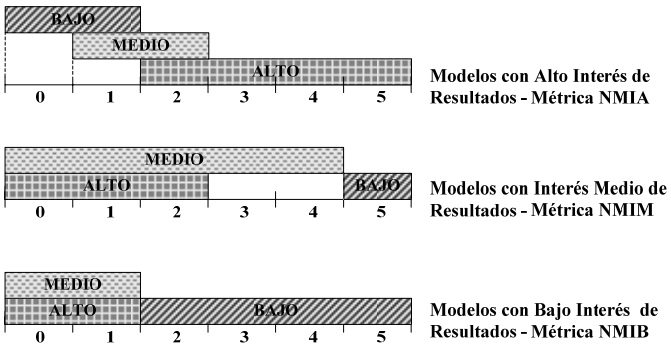

Fig. 34. Distribución de proyectos con alto interés de resultados del modelo para la métrica EPD (P) = Bajo

Se ha planteado una primera definición de una métrica para medir el éxito del proceso de desarrollo en un proyecto de explotación de información para PyMEs, en función de los resultados obtenidos de las pruebas realizadas a los modelos, del conocimiento descubierto y de la utilidad que presenta el mismo para el usuario. La métrica y sus parámetros iniciales propuestos pueden ser refinados para obtener una mayor precisión en su comportamiento. No obstante, se dispone de una primera definición con indicadores de referencia, que puede ser utilizada como forma de medir el éxito logrado por el grupo de desarrollo del proyecto de explotación de información.

4. Métrica de Desvío del Esfuerzo del Proyecto

Para estudiar la métrica de Desvío del Esfuerzo del proyecto de explotación de información – DEFZ (P), se toma en consideración aquellas variables experimentales (sub-sección *a*) que influyen en el comportamiento de la misma.

a) Variables Experimentales

Las variables experimentales que se van a generar para el proceso de simulación son las correspondientes a las métricas básicas que influyen en la métrica DEFZ (P), considerada como métrica derivada. La lista de variables experimentales se muestra en la Tabla CXII.

TABLA CXII. VARIABLES EXPERIMENTALES PARA LA MÉTRICA DEFZ (P)

Variable Experimental	Descripción
DEFZ (P)	Desvío en el esfuerzo para desarrollar el proyecto de explotación de información, el cual se define como: $\text{DEFZ(P)} = \frac{\text{EFZE(P)} - \text{EFZR(P)}}{\text{EFZR(P)}}$
EFZE (P)	Esfuerzo estimado [en persona-mes] para desarrollar el proyecto de explotación de información para PyMEs [65].
EFZR (P)	Esfuerzo real [en persona-mes] aplicado para desarrollar el proyecto de explotación de información.

b) Diseño Experimental

Para Para analizar el comportamiento de la métrica enunciada, se utiliza un banco de pruebas simulado de 32 proyectos de explotación de información, definiendo un rango de valores específico para cada una de las variables experimentales independientes y considerando las restricciones indicadas en [65] según el tamaño del proyecto. De esta manera, se generan diferentes combinaciones sujetas a análisis. Los valores simulados para las variables experimentales independientes se muestran en la Tabla CXIII.

TABLA CXIII. VALORES DEFINIDOS PARA VARIABLES EXPERIMENTALES INDEPENDIENTES DE LA MÉTRICA DEFZ (P)

Variable Experimental	Descripción
EFZE (P)	Esfuerzo estimado [en persona-mes] para desarrollar el proyecto de explotación de información para PyMEs, con un rango de valores específico 6, 12, 18 y 23.
EFZR (P)	Esfuerzo real [en persona-mes] aplicado para desarrollar el proyecto de explotación de información, con un rango de valores específico 20, 40, 60, 80, 90, 110, 120 y 150. Este rango de valores se expresa como un porcentaje del esfuerzo estimado.

El resultado y comportamiento de la métrica derivada DEFZ (P) está asociado al error relativo cometido entre el esfuerzo estimado al inicio del proyecto y el esfuerzo real aplicado para realizar las tareas de cada subproceso del proceso de desarrollo [85], cuyas métricas EFZE (P) y EFZR (P) se indican en la Tabla CXIII. Además, en el caso de la métrica EFZR (P), al estar sometida a un proceso de simulación, no se le puede asignar un número a priori de esfuerzo que fue necesario realizar para cada subproceso, ya que no se conoce el número de personas asignadas al mismo ni su complejidad. Por tal motivo, se decide establecer un rango de valores porcentuales con intervalos, definido entre 20 y 150% que represente el porcentaje de esfuerzo real aplicado respecto del estimado al inicio.

c) Ejecución y Resultado Experimental

En esta sección se analiza el comportamiento de la métrica DEFZ (P), utilizando los datos generados por los experimentos ejecutados.

Del análisis experimental, se observa que cuando el esfuerzo estimado para un proyecto, asociado a la métrica EFZE (P), es mayor al esfuerzo real aplicado, asociado a la métrica EFZR (P), el desvío en el esfuerzo del proyecto toma valores positivos. Mientras que, cuando el esfuerzo estimado es menor al esfuerzo real, el desvío en el esfuerzo del proyecto toma valores negativos. En la Fig. 35 se muestra la variación de esfuerzo por cada proyecto analizado. En la Fig. 36 se muestra el desvío de esfuerzo generado en cada proyecto, en función del esfuerzo estimado y real aplicado para el desarrollo de las tareas de cada subproceso.

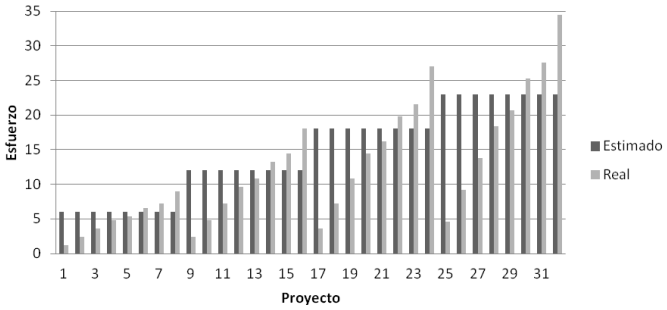

Fig. 35. Esfuerzo estimado y real por proyecto en función de las métricas EFZE (P) y EFZR (P)

d) Regla Experimental

Como regla experimental del comportamiento de la métrica analizada, se concluye que cuando el desvío en el esfuerzo, asociado a la métrica DEFZ (P), toma valores positivos, existe una sobreestimación del esfuerzo necesario para realizar el proyecto de explotación de información. Esto ocurre cuando el esfuerzo estimado es mayor al esfuerzo real requerido para el proyecto. Por el contrario, cuando el desvío en el esfuerzo, asociado a la métrica DEFZ (P), toma valores negativos, existe una subestimación del esfuerzo necesario para realizar el proyecto. Esto ocurre cuando el esfuerzo estimado es menor al esfuerzo real requerido para desarrollar el proyecto.

Una sobreestimación del esfuerzo, puede resultar en la asignación excesiva de recursos al proyecto, una planificación de costos más alta que lo necesario y hasta incluso la cancelación del mismo [74]. La subestimación del esfuerzo, en cambio, puede traer como consecuencia que el proyecto se exceda del presupuesto y fechas de finalización programadas de las tareas, afectándose la calidad del proyecto [74].

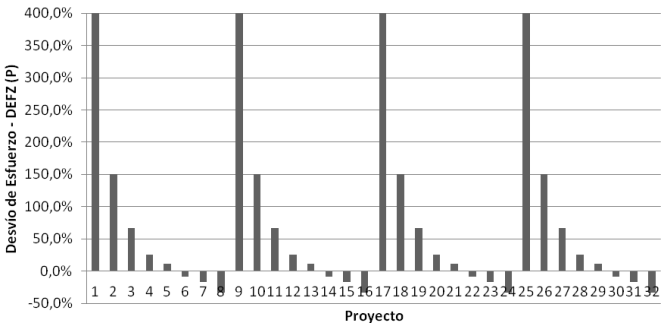

Fig. 36. Desvío del esfuerzo generado por proyecto para la métrica DEFZ (P)

VI. CONCLUSIONES

En este apartado se presentan las aportaciones de este trabajo (sección *A*) y se destacan las futuras líneas de investigación que se consideran de interés en base al problema abierto que se presenta en este trabajo de investigación (sección *B*).

A. Aportaciones de la Investigación

En la Ingeniería de Software como en la Ingeniería del Conocimiento, existen métricas e indicadores que permiten evaluar y analizar distintos atributos y características de los productos y procesos, asegurando la calidad del software producido. Los proyectos de Explotación de Información, también necesitan métricas para medir y controlar el avance del proyecto, evaluar el esfuerzo aplicado, la calidad del proceso y productos desarrollados, y el cumplimiento de los criterios de éxito. Sin embargo, hasta el momento no se disponían de métricas significativas para este tipo de proyectos, ya que por sus diferencias con los proyectos de desarrollo de la Ingeniería de Software e Ingeniería del Conocimiento, las métricas usuales no se consideraban del todo aplicables.

Esta investigación formula aportaciones al cuerpo de conocimiento de la Ingeniería de Proyectos de Explotación de Información con focalización en empresas PyMEs, brindando además, las respuestas a los interrogantes de investigación realizados en el apartado III - sección *C*.

Son contribuciones de esta investigación al cuerpo de conocimiento:

[i]. La definición de tres categorías generales de métricas aplicables a proyectos de Explotación de Información: Métricas de Datos, Métricas de Modelos y Métricas de Proyectos. Cada una de estas categorías incluyen las métricas que cubren el Modelo de Proceso de Desarrollo con las tareas definidas por [85] para este tipo de proyectos: Entendimiento de los Datos, Preparación de los Datos, Modelado (clasificadas para Modelos de Descubrimiento de Grupos, Modelos de Descubrimiento de Reglas y Modelos de Descubrimiento de Dependencias Significativas), Evaluación y Entrega del proyecto.

[ii]. La propuesta de un conjunto de métricas específicas para proyectos de Explotación de Información, tomando como marco de referencia el Modelo de Proceso de Desarrollo definido por [84] para este tipo de proyectos, con particular énfasis en su utilización en empresas PyMEs y proyectos pequeños.

- En la categoría de *Métricas de Datos* se han propuesto las siguientes métricas: Número de Registros, Número de Valores Nulos, Nivel de Compleción de la Tabla, Densidad de Valores Nulos, Número de Valores Erróneos, Grado de Corrección de la Tabla, Número de Atributos No Correctos, Número de Atributos No Significativos, Número de Atributos Sin Errores, Número de Atributos con Defectos, Número de Registros Duplicados, Número de Registros No Correctos, Grado de Utilidad de los Atributos, Número de Atributos Útiles, Número de Atributos No Útiles, Número de Registros Útiles, Número de Registros No Útiles, Número de Atributos Nuevos, Número de Atributos Integrados y Número de Registros Integrados. Estas métricas son aplicables a los subprocesos de Entendimiento de los Datos y Preparación de los Datos para el desarrollo de proyectos de Explotación de Información.
- En la categoría de *Métricas para Modelado* se han propuesto las siguientes métricas: Número de Casos de Entrenamiento del modelo, Número de Casos de Prueba del modelo, Número de Casos de Entrenamiento por clase o grupo, Número de Casos de Prueba por clase o grupo, Número de Clases, Número de Modelos a construir para el proyecto, Número de Atributos a utilizar en el modelo, Número Total de Casos del modelo, Número de Grupos, Número de Casos por Grupo, Exactitud del modelo, Precisión del modelo, Tasa de Aciertos del modelo, Tasa de Errores del modelo, Número de Casos Clasificados Correctamente en su clase o grupo, Número de Casos Clasificados Incorrectamente en su clase o grupo, Precisión de una Regla, Cobertura de una Regla, Número de Casos que satisfacen la precondición de una regla, Número de Casos que satisfacen el consecuente de una regla, Número de Casos Cubiertos por una clase o grupo del atributo clase, Número de Valores Distintos del atributo clase y Grado de Incidencia de un Atributo sobre el atributo clase. Estas métricas son aplicables al subproceso de Modelado para el desarrollo de proyectos de Explotación de Información. A su vez, son compatibles con los procesos de explotación de información definidos por [5], que utilizan tecnologías de sistemas inteligentes.
- En la categoría de *Métricas para Proyectos* se han propuesto las siguientes métricas: Número de Modelos con Alto interés de resultados obtenidos, Número de Modelos con interés Medio de resultados obtenidos, Número de Modelos con Bajo interés de resultados obtenidos y Éxito del Proceso de Desarrollo. Estas métricas son aplicables al subproceso de Evaluación y Entrega para el desarrollo de proyectos de Explotación de Información. Por otra parte, la métrica que mide el Éxito del Proceso de Desarrollo, se relaciona con el cumplimiento de los objetivos de explotación de información, definidos al inicio del proyecto, en el subproceso Entendimiento del Negocio.

[iii]. La utilización de métricas existentes de la Ingeniería de Software, dándoles una interpretación de su uso en los proyectos de Explotación de Información

- Para la categoría de *Métricas de Datos* se han propuesto las siguientes métricas: Número de Tablas y Número de Atributos de la tabla.
- Para la categoría de *Métricas de Proyectos* se han propuesto las siguientes métricas: Duración Real del proyecto, Esfuerzo Real del proyecto, Desvío del Esfuerzo del proyecto, Progreso del proyecto.

[iv]. La utilización de métricas existentes de la Ingeniería del Conocimiento para la categoría de *Métricas de Modelos*, dándoles una interpretación de su uso en los proyectos de Explotación de Información. En esta categoría se han propuesto las siguientes métricas: Número de Reglas descubiertas por el modelo por cada clase o grupo, Número de Atributos de la Precondición de las Reglas de pertenencia a una clase o grupo y Usabilidad de los Atributos del Modelo.

[v]. El estudio del comportamiento de las métricas propuestas para proyectos de Explotación de Información de manera analítica, utilizando simulación por Monte Carlo, y centrado en las particularidades de las empresas PyMEs.

B. Futuras Líneas de Investigación

Durante el desarrollo de esta investigación se han identificado problemas abiertos que por su interés para el campo de conocimiento de los proyectos de Explotación de Información, dan lugar a las siguientes futuras líneas de investigación en la disciplina:

[i]. Durante el desarrollo de la solución se definieron métricas derivadas, que utilizan rangos con valores lingüísticos del tipo ALTO, MEDIO y BAJO. Las conclusiones obtenidas del estudio del comportamiento de estas métricas se basaron en los valores numéricos asignados a estos términos lingüísticos, en base a la propia experiencia. En este contexto surge como problema abierto de interés estudiar si ajustando estos valores numéricos, se obtiene un comportamiento más estable de las métricas derivadas cuando caen en los rangos ALTO, MEDIO y BAJO.

[ii]. El estudio del comportamiento de las métricas propuestas en este trabajo de investigación, se realizó a través de un método empírico de simulación, basado en casos de estudio de complejidad creciente, utilizando el método de Monte Carlo. Con algunas métricas derivadas, se observaron comportamientos un poco imprecisos, ya que dependían de los valores utilizados en la simulación. En este contexto surgen como problemas abiertos de interés:

- Validar las métricas propuestas de manera empírica, 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.
- Validar y ajustar los parámetros de las métricas de manera empírica, utilizando proyectos de Explotación de Información reales, de manera de evaluar la viabilidad de las mismas.

[iii]. El esfuerzo real aplicado al desarrollo de un proyecto de Explotación de Información, se obtiene en base a la cantidad de personas que participaron en cada fase, la experiencia y productividad del equipo de desarrollo y la complejidad de cada fase del proyecto, entre otros factores. Estudiar el comportamiento de la métrica por simulación del método

Monte Carlo, resultó complicado por los factores que debían considerarse. Por tal motivo, se la consideró una métrica básica y no una métrica derivada. En este contexto surge como problema abierto plantear una métrica derivada de esfuerzo real que contemple los factores previamente considerados, para compararla con el esfuerzo estimado al inicio del proyecto.

[iv]. Se ha definido una primera propuesta de métricas que pueden ser utilizadas en proyectos de Explotación de Información, considerando características propias de las empresas PyMEs. En este contexto surgen como problemas abiertos de interés:

- Proponer nuevas métricas de Datos, Modelos y Proyectos que sean aplicables a este tipo de proyectos.
- Ajustar los parámetros de las métricas de manera que sean escalables a proyectos de Explotación de Información de mayor tamaño.

VII. Referencias

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

[2] Agrawal R., Imielinski T., Swami A. 1993. *Mining Association Rules between Sets of Items in Large Databases*. Pp.1-10. ACM SIGMOD International Conference on Management of Data. Washington.

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

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

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

[6] Britos, P., Felgaer, P., García-Martínez, R. 2008c. *Bayesian Networks Optimization Based on Induction Learning Techniques*. In Artificial Intelligence in Theory and Practice II, ed. M. Bramer, (Boston: Springer), 276: 439-443.

[7] Britos, P., Jiménez Rey, E., García-Martínez, E. 2008e. *Work in Progress: Programming Misunderstandings Discovering Process Based On Intelligent Data Mining Tools*. Proceedings 38th ASEE/IEEE Frontiers in Education Conference.

[8] Caruana, R., Niculescu-Mizil, A. 2006. *An Empirical Comparison of Supervised Learning Algorithms*. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA., pp. 161-168.

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

[10] Chen M.S., Han J., Yu P.S. 1996. *Data Mining: An Overview from Database perspective*. IEEE Transactions on knowledge anda Data Engineering, 8(6), pp. 866-883.

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

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

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

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

[15] Fenton, N., Pfleeger S. 1997. *Software Metrics: A Rigorous Approach*. Londres, Chapman & Hall.

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

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

[18] Freitas A. 1999. *On Rule Interestingness Measures*. Elsevier. Knowledge-Based Systems Journal. Volume 12. pp. 309–315.

[19] Freitas A. (2002). *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag. New York. ISBN: 3540433317.

[20] Friedman N., Geiger D., Goldszmidt S. 1997. *Bayesian Network s classifiers*. Machine Learning, 29. Pp 131-163. Kluwer Academic Publishers. The Netherlands.

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

[22] García Jiménez, V. 2010. *Distribuciones de clases no balanceadas: Métricas, Análisis de Complejidad y Algoritmos de Aprendizaje*. Tesis Doctoral. Universitat Jaumé I. Departamento de Lenguajes y Sistemas Informáticos. Castellón. España.

[23] García-Martínez, R., Servente, M., Pasquini, D. 2003. *Sistemas Inteligentes*. Editorial Nueva Librería. Buenos Aires.

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

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

[26] Guyon, I., Elisseeff, A. 2003. *An Introduction to Variable and Feature Selection.* Journal of Machine Learning Research, pp. 1157-1182.

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

[28] Heckerman, D., Chickering, M., Geiger, D. 1995. *Learning bayesian networks, the combination of knowledge and statistical data*. Machine learning 20: 197-243.

[29] IEEE. 1993. IEEE Standard Glossary of Software Engineering Terminology.

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

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

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

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

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

[35] Kalos, M.H., Whitlock P.A. 2008. *Monte Carlo Methods*. Second Edition. John Wiley. ISBN 978-3-527-40760-6.

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

[37] Kan, S. 2002. *Metrics and Models in Software Quality Engineering.* Second Edition. Addison Wesley. ISBN: 0-201-72915-6. Chapter 4.

[38] Klösgen W., Zytkow J.M. 2002. *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press.

[39] Kogan Zahavi, A. 2007. *Integración de Algoritmos de Inducción y Agrupamiento. Estudio del Comportamiento*. Tesis de Grado en Ingeniería en Informática. Universidad de Buenos Aires (UBA). Facultad de Ingeniería. Argentina.

[40] Kohavi, R., Provost, F. 1998. *On Applied Research in Machine Learning*. In Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. Volume 30. Columbia University, New York.

[41] Kohonen, T. 1995. *Self-Organizing Maps*. Springer Verlag Publishers.

[42] Langley P., Sage S. 2013. *Induction of Selective Bayesian Classifiers*. Institute for the Study of Learning and Expertise. Palo Alto, CA-USA. Pp. 399-406.

[43] Larose, D. 2005. Discovering Knowledge in Data, an introduction to Data Mining. John Wiley & Sons. EEUU.

[44] Lavrac N., Flach P., Zupan B. 1999. *Rule Evaluation Measures: A Unifying View*. ILP-99, LNAI 1634, pp. 174-185. Springer-Verlag Berlin Heidelberg.

[45] Lavrac, N., Kavsek, B., Flach, P., Todorovski, L. 2004. *Subgroup discovery with CN2-SD*. Journal of Machine Learning Research, 5. Pp. 153-188.

[46] Liu, B., Hsu W., Chen S., Ma, Y. 2000. *Analyzing the Subjective Interestingness of Association Rules*. IEEE Intelligent Systems, 15(5):47-55.

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

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

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

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

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

[52] McDermid, J. 1991. *Software Engineer's Reference Book*. Editorial Butterworth-Heinemann Ltd.

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

[54] Moser R., Janes A., Russo B., Sillitti A., Succi G. 2005. *Promp: Taking an echography of your software process*. XLIII Congresso Annuale AICA. AGILE Publications. Udine, Italy.

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

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

[57] Olmo Ortíz, J. 2013. *Minería de Datos mediante Programación Automática con Colonia de Hormigas*. Tesis Doctoral. Universidad de Córdoba. Escuela Politécnica Superior. Departamento de Informática y Análisis Numérico. Argentina.

[58] Pérez Cárcamo, P. 2010. *Evaluación de Reglas de Asociación en Text Mining Utilizando Métricas Semánticas y Estructurales*. Tesis de Magister en Ciencias de la Computación. Universidad de Concepción. Facultad de Ingeniería. Chile.

[59] Pfleeger, S. L. 1997. *Assessing Software Measurement*. IEEE Software March/April, pp. 25-26.

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

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

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

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

[64] Pyle, D. 2003. *Business Modeling and Business intelligence*. Morgan Kaufmann Publishers.

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

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

[67] Pytel, P., Amatriain, H., Britos, P., García-Martínez, R. 2012a. 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. Pág. 63-70. Sello Editorial de la Pontificia Universidad Católica del Perú. ISBN 978-612-4057-85-4.

[68] Quinlan, J. 1986. *Induction of decision trees*. Machine Learning, 1(1): 81-106.

[69] Quinlan, J. 1993. C4.5: *Programs for machine learning*. Morgan Kaufmann.

[70] Quinlan, J. R. 2000. C5.0: An Informal Tutorial. Sidney.

[71] Rodríguez D., Pollo- Cattaneo, F., Britos, P., García-Martínez, R. 2010. *Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información*. Anales del XVI Congreso Argentino de Ciencias de la Computación. Pp. 664-673. ISBN 978-950-9474-49-9.

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

[73] Romero Morales, C. 2003. *Aplicación de Técnicas de Adquisición de Conocimiento para la Mejora de Cursos Hipermedia Adaptativos basados en Web*. Tesis Doctoral.

Universidad de Granada. Departamento de Ciencias de la Computación e Inteligencia Artificial. España.

[74] Salvetto de León, P. 2006. *Modelos Automatizables de Estimación muy Temprana del Tiempo y Esfuerzo de Desarrollo de Sistemas de Información*. Tesis Doctoral. Universidad Politécnica de Madrid. Facultad de Ingeniería. Departamento de Lenguajes y Sistemas Informáticos e Ingeniería de Software. España.

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

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

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

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

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

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

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

[82] Tan P., Kumar V. 2000. *Interesting Measures for Association Patterns: A Perspectiva*. Technical Report TR00-036. University of Minnnesota. Department of Computer Science.

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

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

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

[86] Witten, I. H., Frank, E., Hall, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Morgan Kaufmann.

[87] Zhang, C., Fang, Z. 2013. *An Improved K-means Clustering Algorithm*. Journal of Information & Computational Science 10: 1, pp. 193–199.

Diego M. Basso. Es Ingeniero en Informática recibido en la Universidad Nacional de La Matanza y Profesor de Autómatas y Lenguajes Formales y Métricas de Software en la misma Carrera y Universidad. Es Candidato del Programa de Maestría en Ingeniería de Sistemas de Información en la Escuela de Posgrado de la Universidad Tecnológica Nacional (FRBA). Es Investigador Tesista del Laboratorio de Investigación y Desarrollo en Ingeniería de Explotación de Información del Grupo de Investigacion en Sistemas de Información de la Universidad Nacional de Lanús (Argentina).