UNIVERSIDAD NACIONAL DE LA PLATA
FACULTAD DE INFORMÁTICA

Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo

Autor: Ing. Juan Miguel Moine

Directora: Dra. Ana Silvia Haedo (UBA)
Co-directora: Dra. Silvia Gordillo (UNLP)

Tesis presentada en la Facultad de Informática de la Universidad Nacional de La Plata para la obtención del título de Magíster en Ingeniería de Software

Abril de 2013

Dedicatoria:

A todas las personas queridas que me han brindado
su apoyo y motivación durante este camino.

A la Est. Mónica Grasso, quien abrió nuevas puertas en mi vida.

Agradecimiento:

Un profundo agradecimiento a la Dra. Ana Silvia Haedo
y a la Dra. Silvia Gordillo por haberme guiado en este trabajo.

"En Dios confío, los demás traigan datos..."

William Edwards Deming

Resumen

Para llevar a cabo en forma sistemática el proceso de descubrimiento de conocimiento en bases de datos, conocido como minería de datos, es necesaria la implementación de una metodología.

Actualmente las metodologías para minería de datos se encuentran en etapas tempranas de madurez, aunque algunas como CRISP-DM ya están siendo utilizadas exitosamente por los equipos de trabajo para la gestión de sus proyectos.

En este trabajo se establece un análisis comparativo entre las metodologías de minería de datos más difundidas en la actualidad. Para lograr dicha tarea, y como aporte de esta tesis, se ha propuesto un marco comparativo que explicita las características que se deberían tener en cuenta al momento de efectuar esta confrontación.

Contenido

1. Introducción

El descubrimiento de conocimiento en bases de datos, conocido en la actualidad como "minería de datos", es una disciplina que ha crecido enormemente en los últimos años. Las organizaciones han comprendido que los grandes volúmenes de datos que residen en sus sistemas pueden ser analizados y explotados para obtener nuevo conocimiento a partir de los mismos.

Minería de Datos o Explotación de Información[1], es el proceso de extraer conocimiento útil, comprensible y novedoso de grandes volúmenes de datos, siendo su principal objetivo encontrar información oculta o implícita que no es posible obtener mediante métodos estadísticos convencionales.

La entrada al proceso de minería está formada generalmente por registros provenientes de bases de datos operacionales o bien bodegas de datos (Datawarehouse).

El resultado del proceso es un conjunto de patrones (modelos), los cuales serán convertidos en información valiosa para la toma de decisiones.

Los proyectos de minería de datos pueden ser llevados a cabo en distintos escenarios. En el ámbito de las empresas y organizaciones, el más habitual es aquel donde se aborda una situación organizacional (un problema o una oportunidad), buscando patrones y relaciones que puedan colaborar con la misma. Otro tipo de escenario es aquel donde el proyecto comienza con un conjunto de datos y el objetivo es explorarlos para encontrar relaciones interesantes que puedan ser útiles en el dominio de aplicación.

El desarrollo de un proyecto de explotación de información puede dividirse en las siguientes fases:

- **Análisis del negocio**. En esta fase se realiza un análisis de los objetivos de la organización y del problema que se abordará. Se definen los requerimientos del proyecto y se construye un plan de trabajo.
- **Selección, limpieza y transformación de los datos**. Es el conjunto de tareas que tiene por objetivo:
    - Determinar cuáles son las fuentes de información útiles e integrarlas.
    - Depurar y limpiar los datos (por ejemplo hacer un tratamiento de valores atípicos).

[1] En el presente trabajo nos referiremos a los términos "Minería de datos" y "Explotación de información" como procesos de extracción de conocimiento de bases de datos.

- Transformar los datos disponibles según los objetivos del análisis (cambiando el formato a ciertos campos o calculando variables nuevas a partir de las existentes).

Esta etapa del proyecto es conocida también como "pre-procesamiento de los datos", y su ejecución resulta de vital importancia para obtener patrones de buena calidad [15].

Se estima que el 50% del tiempo del proyecto se destina a esta fase por los errores, redundancias e inconsistencias que existen en los datos organizacionales [7], aunque este valor podría resultar sensiblemente menor dependiendo la envergadura del proyecto [27].

- **Modelado**. Consiste en aplicar las distintas técnicas de minería sobre el conjunto de datos para obtener los modelos (patrones) buscados. Estos modelos pueden ser de dos tipos: predictivos o descriptivos.

  Modelos "predictivos" son aquellos que responden a preguntas sobre datos futuros. Permiten estimar el valor de variables de interés, llamadas variables dependientes, a partir de otras llamadas variables independientes. Por ejemplo, una entidad bancaria que necesita predecir qué clientes que solicitan un crédito no lo devuelven. El banco cuenta con datos históricos correspondientes a los créditos que ha otorgado, información de la persona beneficiaria del mismo, y si el crédito fue devuelto o no. A partir de estos datos, se pueden utilizar distintas técnicas para construir un modelo predictivo que determine si es conveniente o no otorgar un crédito a la persona solicitante.

  Dentro de las técnicas de modelado predictivo encontramos, por ejemplo, la regresión logística, los árboles de decisión, redes neuronales y métodos de vecinos más próximos.

  Los métodos "descriptivos" exploran las propiedades de los datos, proporcionando información sobre las relaciones existentes en los mismos. Por ejemplo, una compañía telefónica que desea identificar clientes con gustos similares, con el objetivo de ofrecer promociones especiales a cada grupo.

  Ejemplos de técnicas descriptivas son el análisis de componentes principales y el análisis de conglomerados (clúster).

- **Evaluación de los resultados obtenidos.** Consiste en la evaluación de los modelos obtenidos en la fase anterior, determinando la precisión de los mismos y su interpretación en el dominio del problema.

- **Implementación y difusión**. En esta etapa se incorpora el "nuevo conocimiento" en la toma de decisiones, o bien se documenta y reporta a las partes interesadas [8].

El proceso de minería de datos es iterativo, ya que es posible que en cualquier momento, debamos retroceder a etapas anteriores (Fig. 1).

Fig. 1. Fases generales del proceso de minería de datos

Los esfuerzos en el área de la minería de datos se han centrado en su gran mayoría en la investigación de *técnicas* para la explotación de información y extracción de patrones (tales como árboles de decisión, análisis de conglomerados y reglas de asociación). Sin embargo, se ha profundizado en menor medida el hecho de cómo ejecutar este proceso hasta obtener el "nuevo conocimiento", es decir, en las *metodologías*.

Las metodologías nos permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Ayudan a las organizaciones a entender el proceso de descubrimiento de conocimiento y proveen una guía para la planificación y ejecución de los proyectos. Una metodología no sólo define las fases de un proceso sino también las tareas que deberían realizarse y cómo llevarlas a cabo.

En los inicios del año 1996, KDD (Knowledge Discovery in Databases) [8] constituyó el primer modelo aceptado en la comunidad científica que estableció las etapas principales de un proyecto de explotación de información. En su versión completa, KDD está formado por nueve etapas, donde la primera es el entendimiento del negocio.

A partir del año 2000, con el gran crecimiento en el área de la minería de datos, surgen tres nuevos modelos que plantean un enfoque sistemático para llevar a cabo el proceso [3]: SEMMA [30], CRISP-DM [5] y Catalyst (conocida como P3TQ) [26].

SEMMA fue desarrollada por el SAS Institute y su nombre es el acrónimo correspondiente a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Evaluación).

CRISP–DM, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de minería de datos según un estudio publicado en el año 2007 por la comunidad KDnuggets (Data Mining Community's Top Resource) (Fig. 2) [14]. Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado,

Evaluación e Implementación. Cada fase se descompone en varias tareas generales de segundo nivel.

La metodología Catalyst, conocida como P3TQ (Product, Place, Price, Time, Quantity), fue propuesta por Dorian Pyle en el año 2003. Esta metodología plantea la formulación de dos modelos: el Modelo de Negocio y el Modelo de Explotación de Información. El Modelo de Negocio, proporciona una guía de pasos para identificar un problema (o la oportunidad del mismo) y los requerimientos reales de la organización. El Modelo de Explotación de Información, proporciona una guía de pasos para la construcción y ejecución de modelos de minería de datos a partir del Modelo de Negocio.

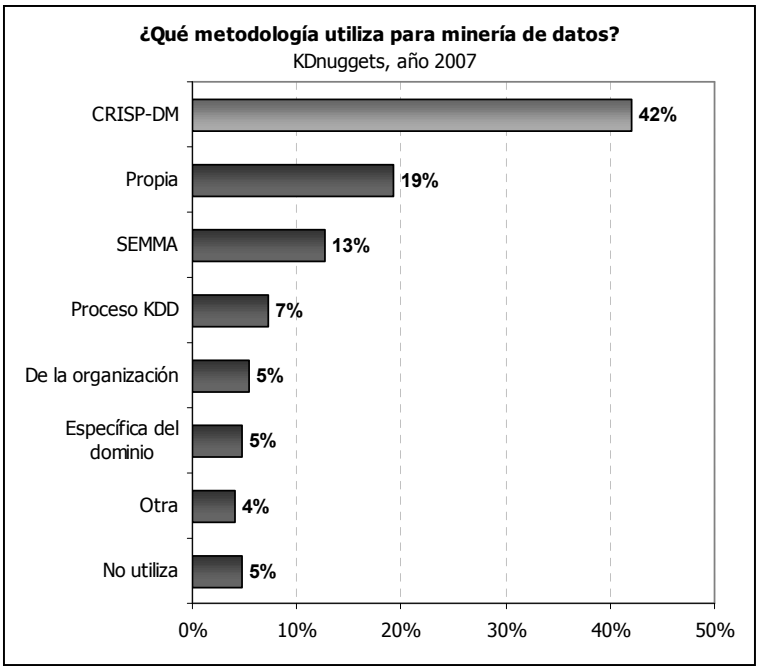

Fig. 2. Encuesta realizada por la KDnuggets en el año 2007

Algunas de las metodologías profundizan en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de minería de datos (como CRISP-DM), mientras que otras proveen sólo una guía general del trabajo a realizar en cada fase (como SEMMA).

En la actualidad son escasos los estudios que comparan las metodologías mencionadas, los cuales están enfocados en establecer un paralelismo entre las fases del proceso [1,16,19] y no un análisis comparativo confrontando las características de cada una.

1.1. Objetivo de la tesis

El objetivo principal de este trabajo de tesis es la construcción de un marco comparativo que sirva como herramienta para poder evaluar y confrontar diferentes metodologías de minería de datos.

El objetivo secundario de esta investigación es el análisis mediante un estudio descriptivo-comparativo de las metodologías más difundidas en la actualidad para proyectos de minería de datos, abriendo una discusión sobre qué enfoques deberían ser realmente considerados una metodología.

1.2. Organización de la tesis

Esta tesis se estructura en seis capítulos.

En el **Capítulo 1** se realiza una introducción global y se plantean los objetivos de este trabajo.

En el **Capítulo 2** se realiza una descripción detallada de los enfoques más difundidos en la actualidad (KDD, CRISP-DM, SEMMA y Catalyst), y se establece una discusión sobre cuáles de ellos pueden considerarse una metodología.

El **Capítulo 3** constituye el eje central de esta tesis, donde se propone un marco comparativo como herramienta para la evaluación y confrontación de metodologías de minería de datos.

En el **Capítulo 4** se analizan en mayor profundidad las metodologías CRISP-DM y Catalyst aplicándolas a un caso de estudio.

En el **Capítulo 5** se confrontan CRISP-DM y Catalyst utilizando el marco comparativo creado en el Capítulo 3.

Finalmente, en el **Capítulo 6** se puntualizan las conclusiones de este trabajo y se establece una discusión sobre futuras líneas de investigación.

1.3. Publicaciones vinculadas a esta tesis

- "Estudio comparativo de Metodologías para Minería de Datos"
  En colaboración con Dra. Ana Silvia Haedo (UBA) y Dra. Silvia Gordillo (UNLP). Trabajo presentado en el "Workshop de Investigadores en Ciencias de la Computación 2011" (WICC). (Rosario, Mayo 2011).

- "Análisis comparativo de Metodologías para la gestión de proyectos de Minería de Datos"
  En colaboración con Dra. Ana Silvia Haedo (UBA) y Dra. Silvia Gordillo (UNLP). Trabajo presentado en el "Workshop de Bases de Datos y Minería de Datos", XVII Congreso Argentino de Ciencias de la Computación 2011 (CACIC). (La Plata, Octubre de 2011).

2. Metodologías para minería de datos

Formalmente, una metodología consiste en un conjunto de actividades organizadas que tienen por objetivo la realización de un trabajo. Para cada actividad se define, además de las entradas y salidas, la forma en la que debe llevarse a cabo.

En este capítulo se analizarán los distintos enfoques para la gestión de proyectos de minería de datos más difundidos en la comunidad científica (KDD, SEMMA, CRISP-DM y Catalyst), aunque existen otros experimentales y de menor difusión que crean híbridos con estándares de Ingeniería de Software [17,20].

2.1. KDD

El Descubrimiento de Conocimiento en Bases de Datos (KDD Knowdlege Discovery in Databases) constituye el primer modelo que define el descubrimiento de conocimiento en bases de datos como un "proceso", compuesto por distintas etapas y fases que van desde la preparación de los datos hasta la interpretación y difusión de los resultados.

En el año 1996, Fayyad define a KDD como el "proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia entendibles en los datos" [11]. El término *proceso* se refiere a la secuencia iterativa de etapas o fases que lo componen. Los patrones deberían ser *válidos* para nuevos datos, *novedosos* en el sentido que deberían aportar nuevo conocimiento al dominio de aplicación y *potencialmente útiles* para el usuario final o tomador de decisiones.

KDD es un proceso iterativo e interactivo. Iterativo ya que la salida de alguna de las fases puede retroceder a pasos anteriores y porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. Es interactivo porque el usuario, o más generalmente un experto en el dominio del problema, debe ayudar a la preparación de los datos y validación del conocimiento extraído.

El modelo de proceso KDD se resume en las siguientes cinco fases (Fig.3):

- *Selección* de los datos sobre los que se trabajará.
- *Pre-procesamiento* de los datos, donde se realiza un tratamiento de los datos incorrectos y ausentes.
- *Transformación* de los datos y reducción de la dimensionalidad.
- *Minería de datos*, donde se obtienen los patrones de interés según la tarea de minería que llevemos a cabo (descriptiva o predictiva).
- *Interpretación y evaluación* del nuevo conocimiento en el dominio de aplicación.

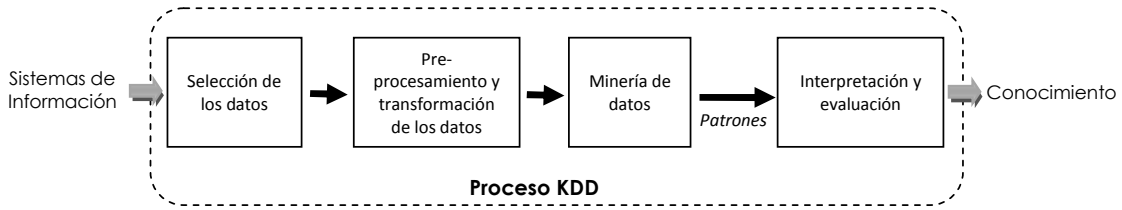

Fig. 3. Esquema del proceso KDD resumido

Formalmente el modelo establece que la minería de datos es la etapa dentro del proceso en la cual se realiza la extracción de patrones a partir de los datos [9,10]. Sin embargo en la actualidad, en la comunidad científica y en la literatura, el término KDD y minería de datos se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento de conocimiento.

Si bien el proceso KDD puede resumirse en las cinco fases mencionadas anteriormente, Fayyad puntualiza nueve etapas para llevarlo a cabo [8]:

1. *Comprensión del dominio de aplicación.* En esta primera etapa, se debería recolectar todo el conocimiento disponible y relevante sobre el dominio de aplicación e identificar los objetivos del proceso KDD desde el punto de vista del usuario.
2. *Creación del conjunto de datos.* Esta etapa consiste en la elección de las fuentes de datos que se utilizarán, la integración de las mismas y la selección de las observaciones/atributos que conformarán la vista minable[2]. Aunque no es estrictamente necesario, en este paso podría requerirse la construcción de un almacén de datos[3].
3. *Limpieza y pre-procesamiento de los datos.* En esta fase se deberían llevar a cabo tareas como limpieza de ruido o datos anómalos (outliers) y tratamiento de datos faltantes (missing values).
4. *Reducción y proyección de los datos.* En este paso se detectan características útiles de representación de los datos dependiendo del objetivo de la tarea de minería (descripción o predicción). Se incluye la utilización de técnicas de reducción de la dimensionalidad y métodos de transformación de los datos para reducir la cantidad de variables en discusión o para encontrar representaciones invariantes de los datos. En esta etapa es frecuente la transformación de los datos, calculando nuevos atributos o bien redefiniendo los existentes con otro formato.

[2] Una *vista minable* es la consolidación en una única tabla de todas las observaciones y los atributos sobre los que se aplicarán los algoritmos de minería.

[3] Almacén de datos, o Datawarehouse, es un repositorio de fuentes heterogéneas de datos, integrados y organizados bajo un esquema unificado para facilitar su análisis y dar soporte a la toma de decisiones.

5. *Determinar la tarea de minería de datos.* En esta fase, se deberá determinar la tarea de minería con la que se abordará el estudio (como agrupamiento, regresión, clasificación, o asociación) teniendo en cuenta los objetivos definidos en la etapa 1.
6. *Determinar el algoritmo de minería.* De acuerdo a la tarea de minería establecida en el punto anterior, en esta etapa se define el algoritmo (o algoritmos) que se aplicarán para la búsqueda de patrones sobre los datos. Incluye la determinación de qué modelos y parámetros son los más adecuados según la naturaleza del problema y de los datos disponibles.
7. *Minería de datos.* Etapa en la que se aplican los algoritmos y técnicas seleccionadas al conjunto de datos en búsqueda de los patrones de interés.
8. *Interpretación.* Comprende la interpretación de los patrones encontrados, visualizando y traduciendo los mismos en términos comprensibles por el usuario.
9. *Utilización del nuevo conocimiento.* En esta fase se implementa el conocimiento descubierto, apoyando con el mismo la toma de decisiones o bien reportándolo a las partes interesadas. Incluye la verificación y resolución de potenciales conflictos con conocimiento descubierto previamente.

Si bien KDD define las fases generales del proceso de minería de datos, no especifica qué actividades puntuales hay que realizar en cada una, quedando la definición de las mismas a criterio del equipo de trabajo.

2.2. SEMMA

SEMMA, creada por SAS Institute, fue propuesta especialmente para trabajar con el software SAS Enterprise Miner [29]. Si bien en la comunidad científica se conoce a SEMMA como una metodología, en el sitio de la empresa SAS se aclara que éste no es el objetivo de la misma, sino más bien la propuesta de una organización lógica de las tareas más importantes del proceso de minería de datos.

SEMMA establece un conjunto de cinco fases para llevar a cabo el proceso de minería (Fig.4): Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado) y Assess (Evaluación). Está especialmente enfocada al desarrollo del modelo de minería, y quedan fuera de su alcance otros aspectos del proyecto como el conocimiento del problema en estudio o la planificación de la implementación.

SAS Enterprise Miner organiza sus herramientas (llamadas "nodos") en base a las distintas fases que componen la metodología. Es decir, el software proporciona un conjunto de herramientas especiales para la etapa de

muestreo, otras para la etapa de exploración, y así sucesivamente. Sin embargo, el usuario podría hacer uso del mismo siguiendo cualquier otra metodología de minería de datos (como CRISP-DM por ejemplo).

Fig. 4. Metodología SEMMA

Las etapas que componen a la metodología son:

1. Sample (Muestreo)

   En esta etapa se toma una muestra del conjunto de datos disponible, que debe ser lo suficientemente grande para contener la información relevante, y lo suficientemente pequeña como para correr el proceso rápidamente. La etapa de muestreo es opcional, aconsejable cuando el tamaño del conjunto de datos es demasiado extenso.

2. Explore (Exploración)

   Consiste en explorar los datos en búsqueda de relaciones y tendencias desconocidas. Es una etapa especial para familiarizarse con los datos, y formular nuevas hipótesis a partir de su análisis.

3. Modify (Modificación)

   Consiste en una etapa de preparación de los datos, donde se limpian los valores anómalos, se realiza un tratamiento de los datos faltantes, y se seleccionan, crean y modifican las variables con las que se trabajarán.

4. Model (Modelado)

   Consiste en la creación del modelo que permitirá predecir las variables de respuesta a partir de las variables explicativas, utilizando algunas de las técnicas predictivas como árboles de decisión, redes neuronales, análisis discriminante o análisis de regresión.

5. Assess (Evaluación)

   En esta fase se evalúa la utilidad y la exactitud de los modelos obtenidos en el proceso de minería de datos, por ejemplo analizando la capacidad predictiva de los mismos.

SEMMA propone que luego de la fase de evaluación, se generan nuevas hipótesis que llevan a repetir el proceso iterativamente (Fig.5).

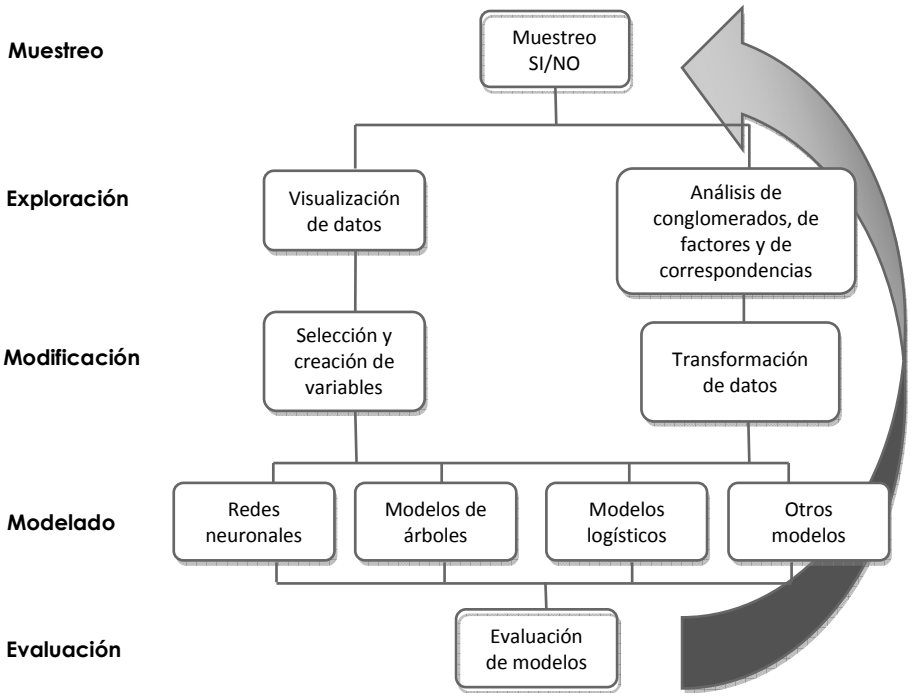

Fig. 5. Iteración de fases en SEMMA

Al igual que en KDD, SEMMA no proporciona una guía de actividades específicas a realizar en cada una de sus etapas. Por este motivo existe una discusión en la literatura acerca de si SEMMA debería ser considerada una metodología.

2.3. CRISP–DM

CRISP-DM (CRoss Industry Standard Process for Data Mining) fue presentada en el año 1999 por las empresas SPSS, Daimer Chrysler y NCR [5,35]. Es una metodología abierta, no está ligada a ningún producto comercial, y fue construida en base a la experiencia de sus creadores, es decir desde un enfoque práctico.

La metodología está estructurada en un proceso jerárquico, compuesto por tareas descriptas en cuatro niveles diferentes de abstracción, que van desde lo general a lo específico.

CRISP-DM, propone en el nivel más alto seis fases para el proceso de minería de datos (Fig.6): entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación e implementación. La sucesión de fases, no es necesariamente rígida.

Cada fase se descompone en un conjunto de tareas genéricas (o generales) de segundo nivel. Estas tareas son genéricas ya que tratan de abarcar la mayoría de las situaciones posibles en minería de datos. A partir del tercer nivel de abstracción, se realiza un "mapeo" de las tareas genéricas definidas

en el modelo a situaciones específicas. De esta forma, las tareas genéricas se traducen en tareas específicas para casos y proyectos concretos. En el cuarto nivel, encontramos las instancias de proceso, donde se describen las acciones, decisiones y resultados de un proyecto particular de minería de datos.

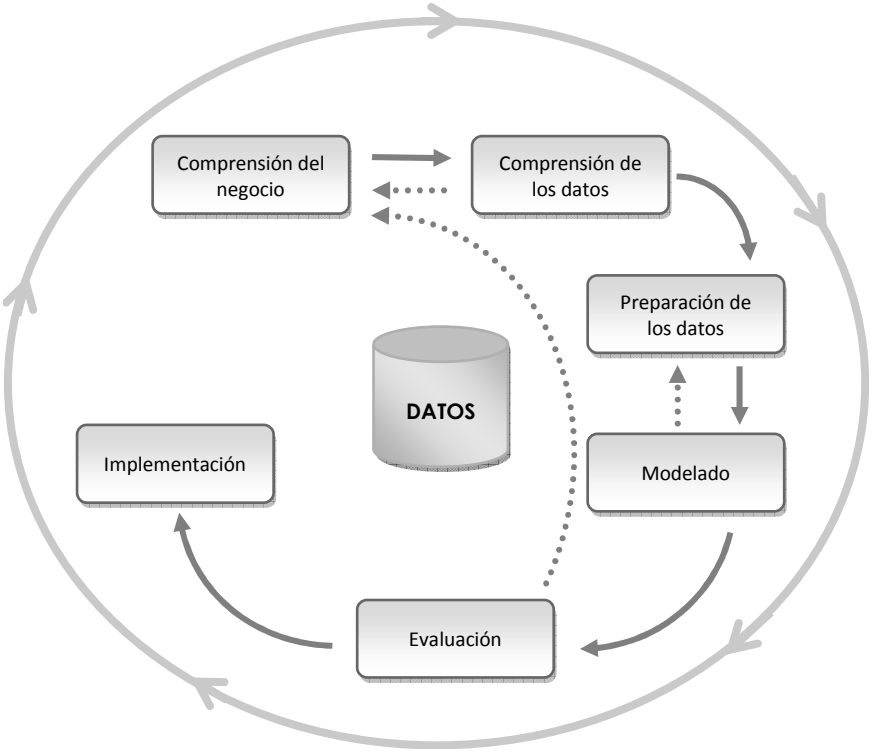

Fig. 6. Metodología CRISP-DM

La metodología proporciona un modelo de referencia y una guía de usuario. El modelo de referencia presenta un resumen de las fases y tareas a llevar a cabo en cada una (junto con sus salidas). Es decir, describe "que" debería hacerse en un proyecto de minería de datos. La guía de usuario proporciona sugerencias para la ejecución de cada tarea del modelo de referencia.

Analizando el nivel más alto de abstracción del modelo, las seis fases que componen el proceso de minería de datos son:

- *Comprensión del negocio:* en esta fase se determinan los objetivos y requerimientos del proyecto desde una perspectiva del negocio, definiendo el problema de minería y el plan de trabajo (Tabla 1).
- *Comprensión de los datos:* fase que consiste en la recolección de datos que se utilizarán en el proyecto y la familiarización con los mismos. En esta etapa es posible el surgimiento de las primeras hipótesis acerca de la información que podría estar oculta (Tabla 2).

- *Preparación de los datos:* comprende aquellas actividades de tratamiento de los datos para construir la vista minable o conjunto de datos final sobre el cual se aplicarán las técnicas de minería (Tabla 3).
- *Modelado:* en esta etapa se aplican las diversas técnicas y algoritmos de minería sobre el conjunto de datos para obtener la información oculta y los patrones implícitos en ellos (Tabla 4).
- *Evaluación:* fase en la que se analizan los patrones obtenidos en función de los objetivos organizacionales. En esta etapa se debería determinar si se ha omitido algún objetivo importante del negocio y si el nuevo conocimiento será implementado, es decir, si se pasará a la próxima etapa (Tabla 5).
- *Implementación:* consiste en la comunicación e implementación del nuevo conocimiento, el cual debe ser representado de forma entendible para el usuario (Tabla 6).

Cada una de estas fases generales se compone de un conjunto de tareas en las que se definen las salidas o entregables que se generan.

Si bien la metodología no especifica detalladamente *cómo* llevar a cabo cada tarea, los consejos de la sección "Guía de Usuario" resultan de mucha utilidad y orientación al momento de ejecutarlas.

Fase 1 CRISP-DM: Entendimiento del Negocio

Tarea	Descripción	Salidas
1.1. Determinar los objetivos del negocio	Entender y establecer cuáles son los objetivos que el cliente pretende alcanzar, desde una perspectiva del negocio.	• Background, con la información que se conoce sobre la situación actual de la organización, incluyendo una descripción general del problema y la solución actual para el mismo (si es que existe). • Objetivos del negocio, identificando los objetivos principales del cliente. • Criterios de éxito, describiendo los resultados esperados desde una perspectiva de negocio.
1.2. Evaluar la situación	Profundizar en la evaluación de la situación actual del negocio. Analizar con mayor profundidad las restricciones y factores que se deben tener en cuenta para el proyecto.	• Inventario de recursos, donde deberán incluirse los recursos disponibles (como los recursos humanos, fuentes de datos, hardware y software). • Lista de requerimientos del proyecto, supuestos y restricciones que se han detectado. • Riesgos y planes de contingencia. Consiste en la identificación de los potenciales riesgos para el proyecto y la planificación de las acciones reactivas que se llevarán a cabo (planes de contingencia). • Glosario con terminología relevante para el proyecto. En el mismo deberá incluirse un glosario de terminología del negocio y otro de minería de datos. • Análisis costo-beneficio del proyecto.
1.3. Determinar los objetivos de la minería de datos	Los objetivos del negocio se describen en términos organizacionales, en cambio los objetivos de minería de datos describen los objetivos del proyecto en "términos técnicos". Es decir, si un objetivo de negocio es aumentar el volumen de ventas, el objetivo de minería de datos podría ser el "agrupamiento" de los clientes para la promoción de nuevas campañas publicitarias.	• Objetivos de minería de datos, describiendo los resultados previstos del proyecto que permiten el logro de los objetivos de negocio. • Definir un criterio de éxito para el proyecto de minería. Especificar las condiciones bajo las cuales se aceptarán los resultados obtenidos.
1.4. Crear un plan para el proyecto de minería de datos.	Crear una planificación para el proyecto de minería, el cual debe ser consistente con los objetivos planteados.	• Plan de proyecto: listar las tareas que deben ser ejecutadas, duraciones y recursos necesarios, así como sus entradas y salidas. El plan del proyecto es un documento dinámico, que debe ser revisado y ajustado al final de cada fase. • Evaluación inicial de técnicas y herramientas de minería que podrían ser utilizadas en el proyecto.

Tabla 1. Entendimiento del negocio en CRISP-DM

Fase 2 CRISP-DM: Entendimiento de los datos

Tarea	Descripción	Salidas
2.1. Recolectar los datos iniciales	Recolectar todos los datos necesarios especificados en la lista de recursos del proyecto.	• Reporte de recolección inicial de datos, donde se detalla la forma en la que han sido obtenidos los conjuntos de datos y los problemas que han surgido en el proceso.
2.2. Describir los datos	Describir en líneas generales los datos recolectados.	• Descripción de los datos, incluyendo el formato de los mismos y su tamaño (como cantidad de registros y variables).
2.3. Explorar los datos	Realizar una exploración de los datos, observando la distribución y comportamiento de las variables con mayor relevancia. En esta fase es conveniente el uso de técnicas simples de análisis estadístico.	• Reporte inicial de exploración de datos, donde se expongan los resultados del análisis y las hipótesis iniciales con su impacto en el proyecto.
2.4. Verificar la calidad de los datos.	Examinar la calidad de los datos, incluyendo un análisis de su completitud, de potenciales errores en los mismos y de los datos ausentes.	• Reporte de calidad de los datos, donde se documente el análisis de calidad efectuado y se propongan potenciales soluciones a los problemas encontrados.

Tabla 2. Entendimiento de los datos en CRISP-DM

Fase 3 CRISP-DM: Preparación de los datos

Tarea	Descripción	Salidas
3.1. Seleccionar los datos	Seleccionar los datos que serán utilizados para el análisis. En esta etapa se debe seleccionar con qué atributos (columnas) y con qué observaciones (filas o registros) se trabajará. La selección debe estar justificada.	• Justificación de la selección. Documento donde se justifiquen las causas por las cuales se incluyeron y excluyeron los datos.
3.2. Limpieza de datos	Es una etapa que tiene por objetivo mejorar la calidad de los datos. En ella se deberán tomar decisiones sobre los problemas de calidad encontrados en los mismos, como datos ausentes o datos anómalos.	• Reporte de limpieza de datos, donde se incluyan las decisiones tomadas sobre los problemas de calidad de los datos (reportados en la fase "2.4 Verificar la calidad de los datos")
3.3.Construcción de los datos	En esta fase se lleva a cabo la construcción de nuevos datos, derivados de los disponibles, que son importantes para el análisis. Estos nuevos datos pueden ser, por ejemplo, atributos calculados o atributos transformados.	• Atributos derivados. Estos atributos se calculan a partir de otros atributos del mismo registro. Por ejemplo: edad_cliente = fecha_venta – fecha_nacimiento. • Registros creados. Estos nuevos registros se crean cuando son necesarios en la fase posterior de modelado.
3.4. Integrar los datos	Consiste en la integración de datos provenientes de diferentes tablas o registros.	• Datos combinados. Resultan de integrar la información de dos o más tablas que tienen diferente información de las mismas observaciones. Por ejemplo, la integración de los datos personales y los datos de las atenciones efectuadas a un paciente en un centro de salud. En esta fase se incluye el cálculo de agregaciones, donde se calculan nuevos datos resumiendo información de diferentes tablas y registros. Siguiendo con el ejemplo del centro de salud, podríamos integrar en un solo registro los datos personales del paciente, el total de atenciones efectuadas, y el promedio anual de consultas médicas realizadas.
3.5. Formatear los datos	Esta etapa se refiere al cambio que debe realizarse en el formato de los datos (pero no en su significado) por los requisitos de las técnicas de modelado elegidas. Por ejemplo, el formato de las fechas o el ordenamiento del set de datos.	• Conjunto de datos reformateados.

Tabla 3. Preparación de los datos en CRISP-DM

Fase 4 CRSIP-DM: Modelado

Tarea	Descripción	Salidas
4.1 Seleccionar la técnica de modelado	Consiste en seleccionar qué técnica de minería de datos será utilizada. Por ejemplo, en un caso donde se ha definido un problema de agrupamiento (clustering), se puede decidir utilizar el algoritmo k-medias. Si se ha optado por el uso de múltiples técnicas, se debería repetir esta tarea para cada una.	• Técnica de modelado. Documentar la técnica de modelado con la que se trabajará. • Supuestos del modelo. Algunas técnicas asumen supuestos sobre el conjunto de datos, como por ejemplo distribución normal de una variable. Documentar todos los supuestos realizados.
4.2. Diseñar las pruebas del modelo	Una vez construidos los modelos, necesitaremos un mecanismo para determinar su calidad y validez. Por ejemplo, en problemas de agrupamiento se puede utilizar el coeficiente de silueta para evaluar la robustez de los grupos encontrados y en problemas de clasificación la tasa de error para estimar la capacidad del clasificador. En esta fase se dividirá el conjunto de datos en un grupo para entrenar el modelo (training) y otro para probarlo (test).	• Diseño de los test. Determinar y documentar de qué forma se entrenarán y evaluarán los modelos generados. Incluir las decisiones tomadas sobre los datos que se utilizarán para entrenamiento y prueba.
4.3. Construir el modelo	Aplicar la técnica seleccionada sobre el conjunto de datos para generar uno o más modelos. En esta fase el modelo será evaluado con distintos valores de parámetros. Por ejemplo, en un algoritmo de agrupamiento k-medias, se podrían generar distintos modelos para diferentes valores de "k" o grupos.	• Parámetros seleccionados. Listar los parámetros que se han proporcionado al modelo, justificando la elección de los mismos. • Modelos producidos por las herramientas de minería. • Descripción de los modelos.
4.4. Evaluar el modelo	En esta fase, el equipo de proyecto interpreta y evalúa el modelo en función de su conocimiento del dominio, los criterios de éxito definidos para el proyecto (tarea 1.3) y las pruebas diseñadas para el modelo (tarea 4.2). Los modelos pueden ser valorados y rankeados.	• Evaluación de los modelos. Generar un reporte de evaluación de los modelos obtenidos, describiendo sus características y un ranking para los mismos. • Evaluación de los parámetros. En función de la evaluación anterior, revisar los parámetros y ajustar los mismos para volver a la fase de construcción del modelo (tarea 4.3). Repetir las etapas 4.3 y 4.4 hasta asegurarse de que se han encontrado los "mejores" modelos.

Tabla 4. Modelado en CRISP-DM

Fase 5 CRISP-DM: Evaluación

Tarea	Descripción	Salidas
5.1. Evaluar los resultados	En esta etapa se evalúa el modelo en función de los objetivos del negocio, determinando su validez de acuerdo a los intereses organizacionales. Además del modelo, puede haber surgido como parte del proceso nueva información relevante y futuras líneas de investigación.	• Evaluación de los resultados de la minería de datos con respecto a los criterios de éxito y objetivos de negocio. • Modelos evaluados y aprobados.
5.2. Revisión del proceso	Realizar una revisión completa del proceso efectuado en búsqueda de posibles errores u omisiones.	• Revisión del proceso, documentando un resumen del mismo. Incluir las actividades omitidas o bien aquellas que deberían ser repetidas.
5.3. Determinar las próximas etapas	En función de la evaluación de resultados y la revisión del proceso, se debe decidir cómo continua el proyecto: si se pasa a la próxima fase (implementación) o bien si se retorna a una fase anterior.	• Lista de posibles acciones. • Descripción de la decisión tomada.

Tabla 5. Evaluación en CRISP-DM

Fase 6 CRISP-DM: Implementación

Tarea	Descripción	Salidas
6.1. Planificar la implementación	En esta etapa se genera el plan de implementación de los resultados obtenidos mediante la minería de datos.	• Plan de implementación, incluyendo las etapas y cómo llevarlas a cabo.
6.2. Planificar el monitoreo y el mantenimiento	El monitoreo y mantenimiento es de gran importancia si los resultados de la minería formarán parte del trabajo diario del negocio y su entorno.	• Plan de mantenimiento y monitoreo.
6.3. Crear un reporte final	Generar un reporte final, que podría resumir el desarrollo del proyecto o bien mostrar un análisis comprensivo de los resultados obtenidos en el proceso de minería.	• Reporte final del proyecto. • Presentación final al cliente, incluyendo resultados y conclusiones.
6.4. Revisión del proyecto	Consiste en identificar y analizar los puntos que fueron bien realizados, los que fueron mal realizados, y los que podrían mejorarse.	• Documentación de la experiencia adquirida durante el desarrollo del proyecto.

Tabla 6. Implementación en CRISP-DM

2.4. Catalyst

En el año 2003, Dorian Pyle propone en su libro "Business modelling and data mining" [26] una metodología para el proceso de extracción de conocimiento en bases de datos llamada "Catalyst". A pesar de ser una metodología muy completa, actualmente no tiene tanto éxito y difusión como CRISP-DM.

Pyle recomienda que el proceso de minería de datos siempre debería colaborar con una situación organizacional, como un problema u oportunidad. Recomienda no trabajar directamente con los datos sino establecer de antemano la problemática que se aborda, el personal involucrado y las expectativas y necesidades de los usuarios. Este punto resulta de gran importancia para justificar la realización del proyecto, ya que difícilmente una organización compre una herramienta si no sabe la función que cumplirá.

Para proyectos donde el problema u oportunidad de negocio no está definido, se recomienda comenzar analizando las relaciones P3TQ - Product (Producto), Place (Lugar), Price (Precio), Time (Tiempo) y Quantity (Cantidad) - que existen en la cadena de valor organizacional. Las relaciones P3TQ se refieren a tener el producto correcto, en el lugar adecuado, en el momento adecuado, en la cantidad correcta y con el precio correcto.

La cadena de valor empresarial (Fig. 7), es un modelo teórico popularizado por Michael Porter que define las actividades de la empresa que van añadiendo valor al producto a medida que éste pasa por cada una de ellas.

Fig. 7. Cadena de valor empresarial

Dentro de los procesos existentes en la cadena de valor, algunos se focalizan en ciertas relaciones P3TQ, como tener el producto apropiado en el lugar correcto, o bien tener la cantidad adecuada en el momento correcto (Fig.8).

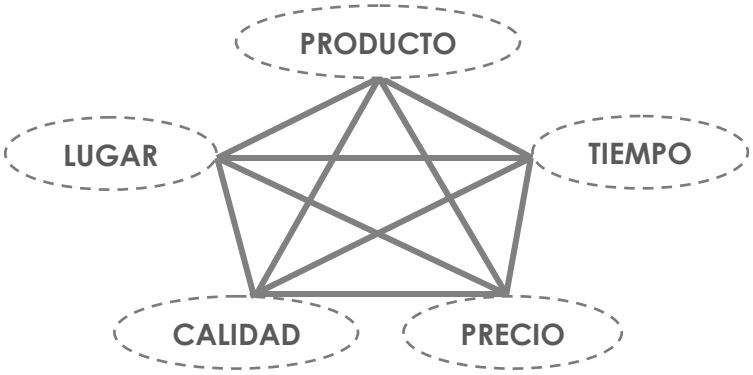

Fig. 8. Relaciones P3TQ

Para proyectos de minería donde se parte de un conjunto de datos y no hay definido un problema de negocio, es importante comenzar identificando el mismo mediante una auditoria de las relaciones P3TQ en la cadena de valor organizacional.

El foco que la metodología "Catalyst" propone sobre la cadena de valor, hizo que la misma sea difundida en la comunidad científica como metodología "P3TQ", aunque esta no sea su denominación original.

2.4.1. Partes de la metodología Catalyst

En cuanto a su estructura, la metodología Catalyst está formada por dos partes (o sub-metodologías): Metodología para el Modelado del Negocio y Metodología para la Minería de Datos.

2.4.1.1. Metodología para el Modelado del Negocio

En esta primera parte se proporciona una guía de pasos para modelar el problema u oportunidad de negocio que abordará el proyecto. Contempla diferentes ámbitos en el proyecto de minería de datos, recomendando una guía de pasos para llevar a cabo en cada escenario específico.

Pyle propone cinco situaciones o puntos de partida diferentes para el proyecto:

- Escenario 1: DATOS

  Explorar los datos en búsqueda de relaciones útiles e interesantes.

  1. Determinar las fuentes de donde se recolectarán los datos.
  2. Identificar al personal interesado (stakeholders) en el proyecto.
  3. Discutir el proyecto original con el personal interesado.
  4. Caracterizar el conjunto de datos en función de las relaciones P3TQ por las cuales fueron recolectados.

5. Caracterizar la motivación del negocio para recolectar y almacenar los datos.
6. Descubrir quién o qué departamento originó el proyecto y qué expectativas tienen sobre el mismo.
7. Descubrimiento del problema
    a. Identificar las principales relaciones P3TQ que dan origen a los datos.
    b. Identificar y caracterizar al personal interesado.
    c. Identificar los objetos organizacionales que los datos representan.
    d. Enmarcar el problema u oportunidad.
    e. Preparar un esbozo del caso de negocio.
    f. Presentar el nuevo proyecto al personal interesado.
    g. Armar el caso de negocio completo, si es necesario.
    h. Enmarcar y describir la situación del negocio.
    i. Definir los requerimientos de la implementación.

- Escenario 2: PROBLEMA/OPORTUNIDAD

  Dado un problema u oportunidad de negocio, ver cómo la minería de datos puede colaborar con la misma.

  1. Identificar y caracterizar al personal interesado relevante.
  2. Explorar la situación de negocio con el personal interesado.
  3. Enmarcar y describir la situación del negocio.
  4. Identificar los objetivos de negocio relevantes para el proyecto.
  5. Buscar los datos que se explotarán.
  6. Armar el caso de negocio.
  7. Presentar el caso de negocio al personal interesado.
  8. Describir la situación del negocio para el proceso de minería.
  9. Definir los requerimientos de implementación.

- Escenario 3: PROSPECCIÓN

  Proyecto diseñado para descubrir dónde la minería de datos puede aportar valor en la organización.

  1. Caracterizar las relaciones P3TQ claves de la organización.
  2. Identificar el flujo de los principales procesos de la organización.
  3. Identificar el personal interesado.

4. Entrevistar al personal interesado.
5. Descubrir qué "cambios estratégicos" pueden resultar de mayor interés para cada usuario.
6. Caracterizar los modelos de minería que pueden dar soporte a los cambios estratégicos.
7. Explorar las fuentes de datos.
8. Preparar un borrador del caso de negocio para cada oportunidad significativa.
9. Presentar los casos de negocio al personal interesado.
10. Enmarcar la situación de negocio que se abordará.
11. Definir los requerimientos de implementación.

- Escenario 4: MODELO DEFINIDO

  Utilizar la minería de datos para construir un modelo específico para una situación determinada.

  1. Identificar al personal interesado.
  2. Discutir los requerimientos con el personal interesado.
  3. Enmarcar la situación del negocio.
  4. Buscar los datos a minar.
  5. Definir los requerimientos de la implementación.

- Escenario 5: ESTRATEGIA

  Dada una situación estratégica, analizar si la minería de datos puede ser útil para explicar la situación actual y descubrir cuáles son las opciones para resolverla. Es decir, el proyecto se inicia requiriendo un análisis estratégico para apoyar la planificación de escenarios corporativos.

  1. Identificar al personal interesado.
  2. Entrevistar al personal interesado.
  3. Enmarcar la situación de negocio.
  4. Si es necesario, trabajar iterativamente con el personal interesado para crear un mapa de los escenarios estratégicos.
  5. A partir del mapa crear un modelo de la situación estratégica, mediante un enfoque de sistema.
  6. Caracterizar las relaciones P3TQ más importantes de la organización.
  7. Relacionar el mapa con las relaciones identificadas.

8. Si es necesario, simular la situación estratégica para identificar ambigüedades, incertidumbres, errores y descubrir relaciones.
9. Caracterizar las relaciones P3TQ en términos de los "cambios estratégicos".
10. Descubrir qué "cambios estratégicos" pueden resultar de mayor interés para cada usuario.
11. Caracterizar los "cambios estratégicos" más viables.
12. Explorar las fuentes de datos.
13. Enmarcar cada oportunidad/problema de negocio en el modelo estratégico con particular atención a las estrategias y sus interacciones, incluyendo los riesgos que implican cada una.

Tomando algunos de estos cinco puntos o escenarios de partida, el autor propone una serie de pasos y herramientas para llegar a descubrir el problema y los requerimientos organizacionales que abordará el proyecto, así como los datos necesarios para efectuar el análisis.

La Figura 9 resume la Metodología para el Modelado del Negocio. En la parte superior se encuentran los posibles escenarios de partida, en el centro las herramientas principales que se pueden utilizar, y en la salida la definición de los datos necesarios y los requerimientos del proyecto.

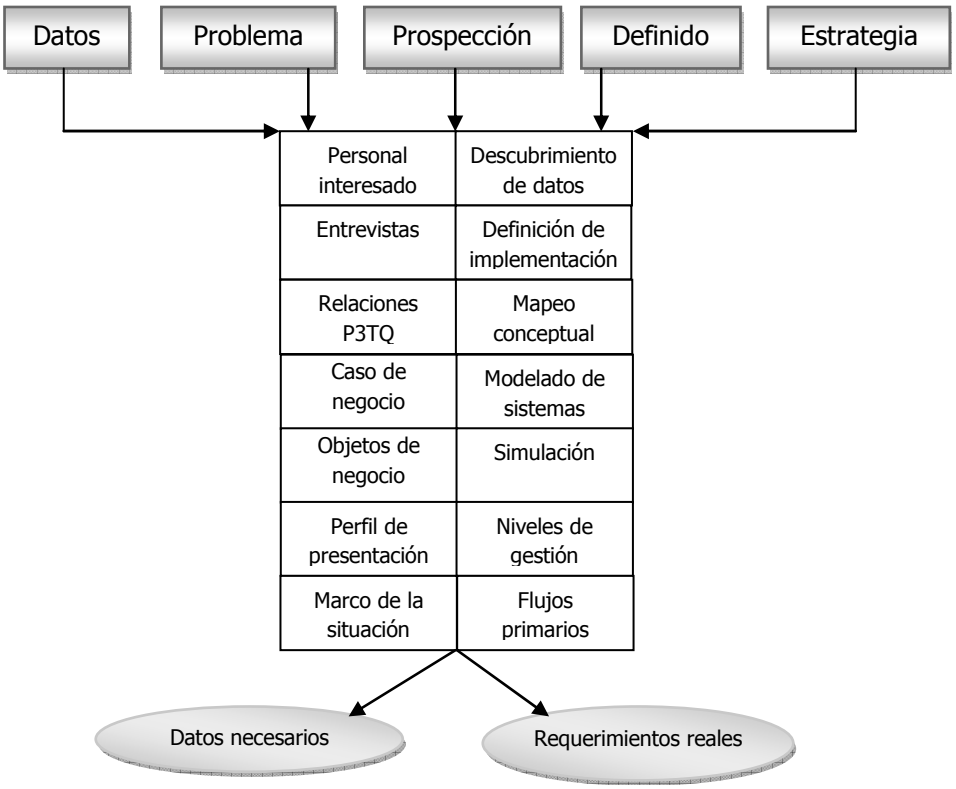

Fig. 9. Metodología de Modelado del Negocio

Cualquiera sea el escenario de partida, siempre se modela el problema que el proyecto de minería abordará, obteniendo:

1. Los datos necesarios para minar.

2. Los requerimientos reales, considerando las expectativas y necesidades del usuario.

2.4.1.2. Metodología para la Minería de los Datos

Proporciona una guía de pasos para el descubrimiento de patrones/relaciones de acuerdo al problema de negocio identificado.

Preparación de los datos

1. Evaluar las variables en estudio (medidas de posición y dispersión, outliers, datos faltantes, etc).
2. Chequear problemas básicos en las variables (variables con muchas categorías por ejemplo).
3. Chequear problemas en la base de datos completa, análisis multivariado (CHAID analysis).
4. Chequear variables anacrónicas (que no aportan información).
5. Chequear que haya suficientes datos.
6. Chequear que se cubran todos los valores posibles de las variables, aun los que no son de interés.
7. Chequear la necesidad de recodificar variables.

Selección de herramientas y modelado inicial

1. Estructurar los datos para el proceso de minería (dividir los datos de entrenamiento, prueba y evaluación).
2. Caracterizar las variables de entrada y salida.
3. Seleccionar el algoritmo de minería.
4. Evaluar el impacto de los datos faltantes mediante un MVCM (Missing Value Check Model).
5. Crear un modelo inicial
    a. Exploratorio / Descriptivo: Si se realiza minería para entender la situación del negocio.
    b. De Clasificación: si se realiza minería para clasificar.
    c. De Predicción: si se realiza minería para predecir.

Refinar el modelo

1. Si el método es exploratorio, describir los resultados encontrados sobre la situación actual.
2. Si el modelo es predictivo o de clasificación, verificar la capacidad predictiva del modelo (por ejemplo con matrices de confusión o gráficos de "valores predichos vs observados", según el caso).
3. Verificar el modelo con el personal interesado.

Implementar el modelo

1. Si el modelo es exploratorio: se deben revisar los requerimientos del problema, elaborar un informe con los resultados del descubrimiento, contabilizar los valores extremos, incorporar evidencia negativa, incorporar evidencia empírica/experimental y obtener realimentación de los usuarios.
2. Si el modelo es de clasificación/predicción: se deben revisar los requerimientos de la implementación planteados antes de la minería, revisar los requerimientos del problema, preparar una explicación del modelo y revisar los requerimientos finales de implementación.
3. Comunicación y difusión de resultados.

2.4.2. Estructuración de la metodología Catalyst

La metodología Catalyst, tanto en la parte de modelado de negocio como en la de minería de datos está compuesta por una serie de pasos llamados "boxes".

La idea es que luego de llevar a cabo una acción, se deben evaluar los resultados y determinar cuál es el próximo paso a seguir (el siguiente "box"). Algunos pasos permiten al modelador elegir entre múltiples caminos dada una situación.

La secuencia y la interacción entre los distintos pasos permiten una flexibilidad muy grande, y una amplia variedad de caminos posibles.

Existen cuatro tipos de "boxes": "Action Boxes", "Discovery Boxes", "Technique Boxes", y "Example Boxes". Cada tipo cumple un rol específico dentro de la metodología.

- Action boxes (AB), en las que se determina cuáles son los siguientes pasos a llevar a cabo en una determinada situación.
- Discovery boxes (DB), en las que se evalúan los resultados y posibles problemas luego de ejecutar una acción.
- Technique boxes (TB), describe cómo utilizar una determinada técnica.
- Example boxes (EB), ejemplifican la aplicación de una técnica. Existe sólo uno en toda la metodología.

En la versión digital de la metodología, todos los pasos se encuentran vinculados mediante hipervínculos para facilitar su navegación.

En la Figura 10 se muestra un ejemplo de Action Box. En este caso, dada la situación de comenzar el proyecto abordando un problema/oportunidad, la metodología lista los pasos a seguir. Cada paso lleva a otro "box".

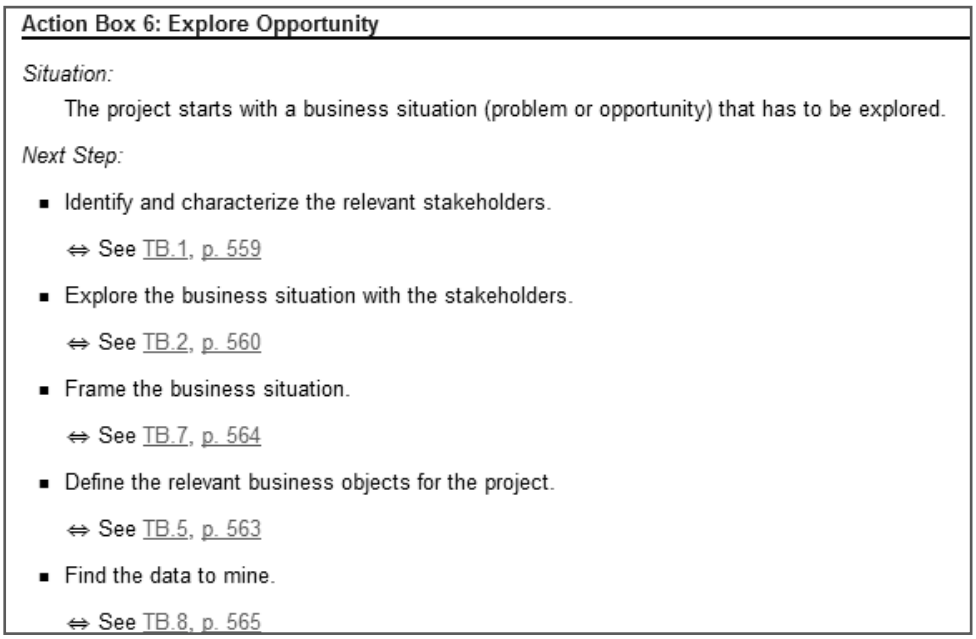

Fig. 10. Action Box de la metodología Catalyst

A través de los "action boxes" y "discovery boxes" Catalyst especifica que actividades realizar en el proceso de minería, proporcionando también información sobre cómo aplicar las técnicas mediante los "technique boxes".

2.5. Análisis de la estructura de cada enfoque

Tomando como base las fases generales que componen a un proceso de minería de datos (análisis del negocio, selección y preparación de los datos, modelado, evaluación, implementación) en la Tabla 7 se establece una correspondencia entre las etapas que componen a los diferentes enfoques.

KDD, CRISP-DM y Catalyst contemplan el análisis y comprensión del problema antes de comenzar el proceso de minería. SEMMA excluye esta actividad del modelo.

En todos los modelos se contempla la selección y preparación de los datos. SEMMA propone trabajar con un muestreo de los datos originales (en caso de tener un gran volumen de datos).

La fase de modelado constituye el núcleo del proceso y está presente en todas las metodologías. El resultado de esta fase es un conjunto de patrones. En SEMMA, la evaluación e interpretación de estos patrones se

realiza sólo sobre el desempeño de los mismos (por ejemplo, la tasa de error), mientras que en las otras metodologías la evaluación se realiza también en función de la utilidad que se aporta al dominio de aplicación o problema organizacional. La implementación de los resultados obtenidos es una fase que no está incluida en SEMMA.

<table>
<tr><th>Fases</th><th>KDD</th><th>CRISP – DM</th><th>SEMMA</th><th>CATALYST</th></tr>
<tr><td>Análisis y comprensión del negocio</td><td>Comprensión del dominio de aplicación</td><td>Comprensión del negocio</td><td></td><td>Modelado del negocio</td></tr>
<tr><td rowspan="3">Selección y preparación de los datos</td><td rowspan="2">Crear el conjunto de datos</td><td rowspan="2">Entendimiento de los datos</td><td>Muestreo</td><td rowspan="3">Preparación de los datos</td></tr>
<tr><td>Comprensión</td></tr>
<tr><td>Limpieza y pre-procesamiento de los datos</td><td rowspan="2">Preparación de los datos</td><td rowspan="2">Modificación</td></tr>
<tr><td>Reducción y proyección de los datos</td></tr>
<tr><td rowspan="3">Modelado</td><td>Determinar la tarea de minería</td><td rowspan="3">Modelado</td><td rowspan="3">Modelado</td><td rowspan="3">Selección de herramientas y modelado inicial</td></tr>
<tr><td>Determinar el algoritmo de minería</td></tr>
<tr><td>Minería de datos</td></tr>
<tr><td>Evaluación</td><td>Interpretación</td><td>Evaluación</td><td>Valoración</td><td>Refinamiento del modelo</td></tr>
<tr><td>Implementación</td><td>Utilización del nuevo conocimiento</td><td>Despliegue</td><td></td><td>Comunicación</td></tr>
</table>

Tabla 7. Correspondencia entre las fases de cada metodología.

2.6. ¿Metodologías o modelos de proceso?

Según Pressman [24], un modelo de proceso es el conjunto de actividades y tareas necesarias para realizar un trabajo, incluyendo la definición de las entradas y salidas de cada una. En una metodología, además de definir para cada actividad las entradas y salidas, se especifican los pasos para su ejecución. Es decir que un modelo de proceso establece "qué hacer", mientras que una metodología define además "cómo hacerlo".

En el desarrollo de este capítulo se ha evidenciado la existencia de dos tipos de enfoques. Por un lado encontramos aquellos que están más cercanos a un modelo de proceso, ya que sólo proponen las actividades generales del proyecto de minería de datos sin explicitar de qué forma se deben llevar a cabo las mismas. Estos modelos son KDD y SEMMA, los cuales no llegan a ser una metodología propiamente dicha y dejan a criterio del equipo de trabajo la definición de las actividades específicas a realizar en cada etapa

del proyecto. Por otro lado, CRISP-DM y Catalyst podrían ser considerados una metodología, porque además de especificar detalladamente las tareas en cada fase proporcionan guías sobre cómo ejecutarlas.

Algunos autores han señalado que CRISP-DM, metodología referente en la comunidad científica, podría completarse aún más si se integrara con herramientas de la Ingeniería de Software, como la elicitación de requerimientos [4]. En algunas investigaciones se han propuesto extensiones de la metodología como en [17,20], donde se integra con estándares internacionales del tipo IEEE 1074, en [21] donde se define un marco para el trabajo colaborativo, o en [22] donde se propone un modelo de proceso ágil.

Si hablamos entonces de metodologías para la gestión de un proyecto de minería de datos, los modelos a tener en cuenta deberían ser CRISP-DM y Catalyst. De aquí en adelante sólo se considerarán como metodologías estos dos enfoques.

3. Un marco comparativo

Una buena gestión de un proyecto de minería de datos requiere la implementación de una metodología de trabajo, como CRISP-DM o Catalyst. A medida que la disciplina madure con el tiempo, irán surgiendo nuevos enfoques y cada vez serán más las alternativas que estarán a disposición del equipo de trabajo.

Surge entonces la necesidad de contar con una herramienta que permita evaluar y confrontar diferentes metodologías. Un marco comparativo que establezca cuáles son los aspectos y características que deben considerarse para seleccionar el enfoque más completo.

El marco comparativo que se desarrolla en este trabajo de tesis está formado por cuatro aspectos (Fig. 11):

1. Nivel de detalle en las actividades de cada fase.

   En este aspecto se evalúa el grado de profundidad con el que se describen las actividades a realizar en cada fase del proceso. Para que un enfoque pueda ser considerado una metodología, debería proponer las actividades específicas que componen cada etapa y una guía de cómo llevarlas a cabo.

2. Escenarios de aplicación.

   Se refiere a cómo la metodología se adapta a los diferentes escenarios que pueden constituir el punto de partida del proceso.

3. Actividades específicas que componen cada fase.

   En este aspecto se evalúan las actividades principales que deberían estar presentes en cada fase del proceso de minería de datos.

4. Actividades destinadas a la dirección del proyecto.

   Las actividades de dirección del proyecto están enfocadas en la administración de ciertos aspectos tales como el tiempo, el costo o el riesgo. En esta sección se evalúa la incorporación de estas actividades que tienen por objetivo aumentar las probabilidades de que el proyecto finalice exitosamente en el tiempo estimado con el presupuesto aprobado.

Para cada uno de estos aspectos se propone la evaluación de una serie de características, las cuales deberían estar presentes en una metodología de minería de datos bien definida.

Fig. 11. Aspectos del marco comparativo.

3.1. Aspecto 1: Nivel de detalle en las actividades de cada fase.

Una metodología está formada por un conjunto de actividades, las cuales por lo general son agrupadas en un mayor nivel de abstracción denominado fase. La cantidad de niveles de abstracción puede variar entre una metodología y otra. Lo importante es que para cada fase, se definan un conjunto de actividades específicas que secuencien el trabajo a realizar. En una metodología completa, se espera que no sólo se describan las actividades, sino también se especifique la forma en la que deben llevarse a cabo. El resultado tangible de una actividad se denomina entregable.

En este aspecto del marco comparativo se evaluará el nivel de detalle con el que una metodología define las actividades que conforman al proceso.

- *Característica 1.1. ¿Se definen actividades específicas para cada fase del proceso?*

  Para cada fase que compone el proyecto, una metodología de minería de datos debería proponer un conjunto de actividades específicas que detallan el trabajo a realizar.

  Resultado esperado: Para cada una de las fases del proceso se definen actividades específicas de menor nivel.

- *Característica 1.2. ¿Se explicitan los pasos a seguir para llevar a cabo cada actividad?*

  Para reducir la subjetividad y dar mayor apoyo al usuario durante el proceso, una metodología completa debería describir los pasos necesarios para ejecutar cada una de las actividades específicas.

  Resultado esperado: Se indica paso a paso cómo ejecutar cada una de las actividades específicas que se proponen.

- *Característica 1.3. ¿Se definen las entradas de cada actividad?*

  La definición de las entradas (inputs) de una actividad le permitirá al equipo de trabajo saber cuándo está en condiciones de comenzar la misma y qué elementos requiere para ello (prerrequisitos). Una entrada puede ser el resultado de una actividad anterior.

  Resultado esperado: Se indican las entradas necesarias para cada actividad.

- *Característica 1.4. ¿Se definen las salidas de cada actividad?*

  La salida de una actividad (output) generalmente se materializa mediante uno o más entregables, los cuales representan los resultados obtenidos luego de su ejecución.

  Resultado esperado: Se especifican las salidas o entregables de cada actividad.

- *Característica 1.5. ¿Se provee una guía de buenas prácticas para cada una de las actividades específicas?*

  Para el usuario de la metodología, especialmente si es principiante, resulta muy importante contar no sólo con la definición de las actividades específicas sino también con una serie de consejos desde el punto de vista práctico que reduzcan los inconvenientes que le pudieran surgir durante su ejecución.

  Resultado esperado: Se proponen consejos prácticos para tener en cuenta en la ejecución de cada actividad.

3.2. Aspecto 2: Escenarios de aplicación.

Los proyectos de explotación de información pueden ser llevados a cabo en distintos escenarios. En algunos casos el usuario desea obtener nuevo conocimiento para abordar algún problema/situación, y en otros se encuentra interesado en explorar sus datos transaccionales en busca de relaciones o patrones que puedan serle útiles. Sin embargo, en este último caso también existe una situación de trasfondo que motiva el proyecto, ya que difícilmente una organización lo financie si no se establece los beneficios que producirá.

- *Característica 2.1. ¿Se especifican actividades para la definición y el análisis del problema u oportunidad con el cual colaborará la minería de datos?*

  El punto de partida para el proyecto debería ser el análisis del problema o situación para el cual se desea obtener el "nuevo conocimiento". Esta actividad permitirá definir las fuentes de datos necesarias y las técnicas

de minería más apropiadas para la problemática que se aborda. La definición del problema u oportunidad de negocio no es una tarea sencilla, ya que requiere interiorizarse con la situación del usuario y definir claramente cuáles son los objetivos del proyecto.

Resultado esperado: Se considera en las primeras fases del proceso actividades específicas para definir y analizar el problema u oportunidad organizacional en el que se enmarcará el proyecto.

- *Característica 2.2. ¿Se consideran puntos de partida alternativos donde el usuario no refiere un problema sino que sólo desea explorar sus datos?*

  En ocasiones el usuario manifiesta que desea implementar un proyecto de minería de datos para encontrar patrones ocultos en la información de su organización. Bajo esta situación, el analista estaría navegando por la información buscando cualquier tipo de conocimiento novedoso. En estos casos siempre existe una problemática o interés de trasfondo que motiva a dicha exploración.

  Resultado esperado: En aquellos casos donde aparentemente se desea "explorar" la información organizacional, la metodología provee una guía de actividades para identificar problemas latentes que el usuario desconoce y que permitirán definir objetivos claros para el proyecto.

- *Característica 2.3. ¿La metodología es independiente del dominio de aplicación?*

  Los proyectos de minería de datos pueden llevarse a cabo en diversas áreas, tales como salud, industria, comercio, deporte o educación. Si bien cada dominio requiere de conocimientos específicos, las actividades principales del proceso de explotación de información deberían ser independientes del ámbito de aplicación.

  Resultado esperado: La metodología es general y no está condicionada a un dominio de aplicación en particular.

- *Característica 2.4. ¿La metodología es aplicable a proyectos de diferente tamaño?*

  Al igual que en otras disciplinas, los proyectos de minería de datos pueden ser de baja, mediana o gran envergadura. Una metodología debería poder gestionar proyectos de cualquier tamaño. El equipo de trabajo podrá seleccionar aquellas actividades que crea conveniente según el tamaño de su proyecto.

  Resultado esperado: La metodología puede utilizarse para proyectos de cualquier tamaño.

3.3. Aspecto 3: Actividades específicas que componen cada fase.

En este aspecto se pretende analizar la incorporación de ciertas actividades relevantes que deberían estar presentes a lo largo del proceso de minería de datos. Para ello, se propone la evaluación de una serie de características en función de las distintas fases generales que componen al proceso: análisis del problema, selección y preparación de los datos, modelado, implementación y evaluación.

Fase de análisis del problema

- *Característica 3.1. ¿Se propone una evaluación general de la organización?*

  Antes de comenzar con la definición de las necesidades del usuario, resulta de gran importancia llevar a cabo un relevamiento general de la organización. En este relevamiento se deben incluir los objetivos del negocio, la estructura de la organización y una breve descripción de las actividades que se desarrollan en la misma.

  Resultado esperado: La metodología comienza el proyecto con una evaluación general de la organización.

- *Característica 3.2. ¿Se identifica al personal involucrado en el proyecto (stackeholders)?*

  Consiste en identificar a todas aquellas personas afectadas por el proyecto o que tienen algún tipo de interés sobre el mismo.

  Resultado esperado: Se identifica y caracteriza a las personas involucradas e interesadas en el proyecto.

- *Característica 3.3. ¿Se define el problema u oportunidad de negocio?*

  La definición del problema u oportunidad resulta de vital importancia para definir qué tipo de patrones se buscan e identificar los objetivos del proyecto.

  Resultado esperado: Se proponen actividades específicas que permitan detectar y definir la problemática (o situación) organizacional con la cual colaborará el proyecto de minería de datos.

- *Característica 3.4. ¿Se propone una evaluación de las fuentes de datos?*

  En esta primera fase del proyecto es conveniente llevar a cabo un estudio de las fuentes de datos que se requerirán como entrada para el proceso de explotación de información. Además de la identificación de estas fuentes, es importante la definición de su estructura y accesibilidad

para poder planificar correctamente la construcción del conjunto de datos.

Resultado esperado: La metodología sugiere una evaluación temprana de las fuentes de datos que se requerirán para el proyecto.

- *Característica 3.5. ¿Se analizan todas las soluciones posibles al problema?*

  La minería de datos generalmente no es la única alternativa que puede colaborar en la resolución de un problema. Para mejorar la justificación del proyecto se deberían analizar todas las soluciones posibles al problema, incluyendo la posibilidad de "no hacer nada".

  Resultado esperado: Se definen actividades para analizar las diferentes alternativas de solución al problema y especificar porqué la realización de un proyecto de minería de datos se encuentra dentro de las más factibles.

- *Característica 3.6. ¿Se especifican los objetivos del proyecto?*

  Una vez definido el problema u oportunidad de negocio, se debe especificar qué objetivos tendrá el proyecto desde el punto de vista organizacional y técnico.

  Resultado esperado: Se especifican los objetivos del negocio y los objetivos técnicos del proyecto.

- *Característica 3.7. ¿Se define un criterio de éxito para el proyecto?*

  Consiste en explicitar cuándo se considerará que el proyecto de minería de datos ha logrado resultados aceptables para satisfacer los objetivos planteados.

  Resultado esperado: Se definen los criterios de éxito para el proyecto en el plano organizacional y en el plano técnico.

- *Característica 3.8. ¿Se realiza una evaluación general de las técnicas de minería que podrían utilizarse?*

  Consiste en el análisis de las técnicas de minería que son más adecuadas para llevar a cabo los objetivos del proyecto. Esta tarea es importante para resolver los principales aspectos técnicos en etapas tempranas.

  Resultado esperado: Se realiza una evaluación inicial de las técnicas de minería más adecuadas para el proyecto.

- *Característica 3.9. ¿Se especifica de qué forma el usuario utilizará el nuevo conocimiento?*

Esta característica se refiere a dejar documentada la forma en la que el "nuevo conocimiento" se entregará, difundirá y utilizará. Resulta de gran ayuda para explicar al usuario cómo los patrones descubiertos se incorporarán en los procesos de la organización o bien en la toma de decisiones estratégicas.

Resultado esperado: Se sugiere en etapas tempranas la especificación de la difusión y uso de los patrones obtenidos.

Fase Selección y Preparación de los datos

- *Característica 3.10. ¿Se propone un análisis exploratorio inicial de los datos?*

  Una vez seleccionados y recolectados los datos, es conveniente realizar un análisis exploratorio para familiarizarse con los mismos. Un estudio de la distribución y comportamiento de las variables, identificando además las tareas de limpieza que se deberán llevar a cabo.

  Resultado esperado: Se efectúa un análisis exploratorio/descriptivo inicial de los datos recolectados.

- *Característica 3.11. ¿Se sugieren actividades para la limpieza de los datos?*

  Los datos recolectados de las bases de datos transaccionales generalmente no son perfectos. Algunos campos y observaciones podrían estar incompletos. A su vez, podría ocurrir también la existencia de valores atípicos que deforman la descripción de las variables en estudio. La limpieza de datos será de gran ayuda para descubrir patrones que reflejen mejor la realidad, y no caer en el fenómeno GIGO[4].

  Resultado esperado: Se proponen actividades de limpieza de datos previamente a la construcción de la vista minable.

- *Característica 3.12. ¿Se contemplan actividades para la transformación de variables y la creación de atributos derivados?*

  Generalmente los datos originales no tienen el formato que se necesita para la vista minable, razón por la cual resulta de gran importancia la transformación de las variables existentes y el cálculo de atributos

4 Garbage In – Garbage Out (GIGO). En el ámbito de minería de datos es un término que se emplea para indicar que si los datos de entrada son de mala calidad, los patrones obtenidos serán malos (si entra basura, saldrá basura).

derivados. Los atributos derivados son aquellos cuyo valor surge a partir de otros campos del mismo registro.

Resultado esperado: Se especifican actividades para la transformación de los datos.

- *Característica 3.13. ¿Se realiza un análisis descriptivo final sobre los datos depurados?*

  Una vez que se ha realizado la limpieza de los datos y las transformaciones necesarias sobre los mismos, el conjunto de datos está preparado para la aplicación de las técnicas de minería. Un análisis descriptivo final del conjunto de datos terminado (vista minable) será de gran importancia para una mejor modelización e interpretación durante la fase de modelado.

  Resultado esperado: Se sugiere un análisis descriptivo final sobre el conjunto de datos terminado.

- *Característica 3.14. ¿Se verifica con el usuario la completitud del conjunto de datos final?*

  Para que los patrones resultantes del proceso sean buenos, el conjunto de datos debe representar lo mejor posible la realidad. Es decir, se debe chequear que se contemplen todos los casos posibles para que no queden valores de las variables en estudio sin considerar.

  Resultado esperado: Se verifica la completitud del conjunto de datos con el usuario.

Fase de Modelado

- *Característica 3.15. ¿Se efectúa una selección de las técnicas que se aplicarán?*

  En función de la tarea de minería que se haya planteado (como clasificación o agrupamiento) existen diversas técnicas que el modelador puede utilizar. Sin embargo no todas las técnicas son aplicables en todos los casos. Algunos factores pueden influir en la selección de las mismas, como el tamaño del conjunto de datos o bien la naturaleza de las variables en estudio.

  Luego de analizar detalladamente la factibilidad de cada técnica, se debería efectuar la selección final de aquellas que se utilizarán en esta fase de modelado.

  Resultado esperado: Se efectúa un análisis y selección final de las técnicas de minería que se implementarán para la creación de los modelos.

- *Característica 3.16. ¿Se planifica de qué forma se evaluarán los resultados?*

  Previamente a la aplicación de las técnicas es importante establecer cómo se evaluarán los resultados obtenidos por las mismas. En este punto debería definirse cómo se dividirá el conjunto de datos para el entrenamiento y prueba de los modelos. Por otro lado debería establecerse cuál será el criterio para la ponderación de los resultados.

  Resultado esperado: Se propone, previamente a la aplicación de las técnicas, la planificación de cómo se evaluarán los resultados obtenidos.

- *Característica 3.17. ¿Se efectúa una evaluación inicial de los modelos obtenidos?*

  Una vez obtenidos los modelos de minería, es importante llevar a cabo una descripción de los mismos en función de los criterios de evaluación que se hayan especificado. Esta evaluación inicial estaría centrada en aspectos técnicos (por ejemplo, en problemas de clasificación, calculando para cada modelo la tasa de error). En la fase siguiente se compararán los resultados obtenidos, determinando cuáles son los más adecuados en función de los objetivos planteados.

  Resultado esperado: Se sugiere una evaluación técnica inicial de cada modelo obtenido.

- *Característica 3.18. ¿Se proveen directivas para el caso donde se dificulta el descubrimiento de los patrones?*

  El proceso de modelado no siempre arroja buenos resultados. En este caso, la experiencia del modelador y las actividades alternativas que propone la metodología resultan de gran importancia para no darse por vencido y continuar con el proceso de modelado hasta que se obtengan resultados aceptables.

  Resultado esperado: Se especifican caminos alternativos para el caso donde no se logren descubrir patrones en el conjunto de datos.

Fase de Evaluación

- *Característica 3.19. ¿Se interpretan los modelos en función de los objetivos organizacionales?*

  Además de evaluar la calidad de los modelos obtenidos desde un punto de vista técnico, es necesario analizar la adecuación de los mismos a los objetivos organizacionales. En este análisis deberían interpretarse los resultados en función de la situación del negocio, determinando si los mismos resultan útiles desde una perspectiva organizacional. La

participación del usuario resulta de gran importancia, ya que colaborará en la validación de los patrones descubiertos.

Resultado esperado: Se propone la interpretación de los modelos en función de los objetivos del negocio.

- *Característica 3.20. ¿Se comparan y ponderan los modelos obtenidos?*

  Luego de analizar cada modelo individualmente, es necesario establecer una ponderación para determinar cuáles son los más robustos y que mejor se adecúan a los objetivos planteados.

  Resultado esperado: Se sugiere una comparación entre los modelos obtenidos, para posteriormente establecer una ponderación de los mismos en función de los objetivos técnicos y organizacionales.

- *Característica 3.21. ¿Se propone una revisión general del proceso?*

  Para asegurarse de que ningún punto importante se ha omitido durante el desarrollo del proyecto, se debería proponer la revisión general del proceso, analizando la concordancia entre los objetivos planteados y los resultados obtenidos.

  Resultado esperado: Se propone una revisión general del proceso donde se analiza la consistencia y completitud del mismo.

- *Característica 3.22. ¿Se proveen directivas en caso de que ninguno de los modelos obtenidos resulte viable?*

  En este punto del proyecto puede suceder que luego de realizar un análisis completo de cada modelo, ninguno de ellos resulte viable de implementar ya sea porque no aportan conocimiento novedoso, no son factibles desde el punto de vista técnico o bien su interpretación es deficiente. En este caso la metodología debería proponer actividades para retomar el proceso en fases anteriores, reformular los objetivos del proyecto o bien dar por finalizado el mismo.

  Resultado esperado: Se definen acciones reactivas para el caso donde ninguno de los modelos obtenidos resulta viable de implementar.

Fase de Implementación

- *Característica 3.23. ¿Se planifica la implementación del nuevo conocimiento?*

  Consiste en efectuar un plan donde se especifiquen las actividades que se llevarán a cabo para dar difusión y uso del nuevo conocimiento. En la descripción de estas actividades se debe tener en cuenta si los modelos

serán utilizados para la toma de decisiones o serán incorporados en los procesos operacionales de la organización.

Resultado esperado: Se propone una planificación de las actividades necesarias para implementar el nuevo conocimiento dentro de la organización.

- *Característica 3.24. ¿Se propone la creación de un programa de mantenimiento?*

  Una vez que los modelos fueron implementados, si los mismos son utilizados en los procesos operacionales, se debe comprobar periódicamente que sigan siendo válidos. Esta situación se genera porque probablemente, a lo largo del tiempo, las condiciones del entorno cambien pudiendo afectar a la validez de los modelos.

  Resultado esperado: Se sugiere la creación de un plan de mantenimiento que permita controlar periódicamente la validez de los modelos implementados.

- *Característica 3.25. ¿Se entrega al usuario un resumen del proyecto?*

  La creación de un reporte final es importante para exponer un resumen del proyecto, donde se vinculen los aspectos más importantes del mismo. Es de gran utilidad para comunicar a los usuarios la evolución del proceso y los resultados obtenidos.

  Resultado esperado: Se propone la creación de un reporte con un resumen del proyecto.

- *Característica 3.26. ¿Se documenta la experiencia adquirida por el equipo de trabajo?*

  Consiste en actividades de revisión donde se destacan buenas y malas prácticas efectuadas durante el trascurso del proyecto. Este documento sería como un análisis post-mortem, el cual resulta de gran utilidad para dejar documentada la experiencia adquirida durante el mismo (¡y no volver a cometer los mismos errores!).

  Resultado esperado: Se realiza una revisión de todo el proyecto, documentando la experiencia adquirida durante el mismo.

3.4. Aspecto 4: Actividades de dirección del proyecto

Las actividades de dirección del proyecto consisten en la planificación, ejecución y control de ciertos aspectos importantes para que el mismo se desarrolle exitosamente. Entre estos aspectos se encuentran, por ejemplo, la administración del costo (presupuesto) y la del tiempo del proyecto (cronograma).

Este tipo de actividades se clasifica en dos grupos: actividades de planificación y actividades de control. Las actividades de planificación incluyen la identificación de las tareas a realizar en el proyecto, estimación de la duración de las mismas, estimación de los recursos afectados y la definición del curso de acción. Las actividades de control tienen por objetivo el monitoreo del estado actual del proyecto para su comparación con lo planificado.

Existen ciertos estándares reconocidos internacionalmente, como el PMBOK [25], que establecen cuáles son las áreas que deben administrarse para lograr una adecuada gestión del proyecto, independientemente de su tipo.

Tomando como referencia el PMBOK, en este aspecto del marco comparativo se definen cinco áreas de dirección del proyecto, para cada una de las cuales se evaluarán características relacionadas a la planificación y control de las mismas.

Las áreas que se proponen en este marco comparativo son:

1. Gestión del alcance.
2. Gestión del tiempo.
3. Gestión del costo.
4. Gestión del equipo de trabajo.
5. Gestión del riesgo.

Gestión del alcance

La gestión del alcance consiste en la planificación y control de todo el trabajo que se ejecutará en el transcurso del proyecto. Para ello se procede a la definición de aquellos entregables (o sub-productos) que serán incluidos en el proyecto, delimitando de esta forma el trabajo que se realizará.

Cabe hacer una distinción entre alcance del producto y alcance del proyecto. El primero se refiere a las características que debe tener el producto mientras que el segundo se refiere al trabajo que se llevará a cabo para entregar el producto. En esta área cuando hablamos del alcance nos estaremos refiriendo al último caso.

- *Característica 4.1. ¿Se propone la selección de los entregables que se generarán durante el proyecto?*

  Trabajar siguiendo una metodología no significa que se deban realizar todas las actividades y crear todos los entregables que la misma proponga. La metodología constituye una guía de referencia, cada equipo de trabajo deberá seleccionar cuáles son los entregables que formarán parte de su proyecto.

Una herramienta muy utilizada es la Estructura de Desglose de Trabajo (o Work Breakdown Structure) la cual organiza en forma jerárquica el trabajo que será ejecutado.

Resultado esperado: Se propone la selección y definición de las actividades y entregables que se irán desarrollando a lo largo del proyecto.

- *Característica 4.2. ¿Se especifican actividades de control del alcance?*

  Consiste en detectar los cambios en el trabajo que debe realizarse y mantener actualizado el listado de actividades y entregables que se ha planificado.

  Resultado esperado: Al final de cada fase se propone una revisión de las actividades y entregables planificados para determinar si es necesario algún cambio.

Gestión del tiempo

Esta área está formada por aquellas actividades cuyo objetivo es lograr la conclusión del proyecto en el tiempo estipulado.

- *Característica 4.3. ¿Se realiza una definición y secuenciación de las actividades que se ejecutarán durante el proyecto?*

  Consiste en la definición de las actividades que se ejecutarán en cada fase del proyecto. Una vez definidas las actividades se deben analizar las dependencias y secuenciar la forma en la que se llevarán a cabo.

  Resultado esperado: Se definen qué actividades específicas se ejecutarán en cada etapa del proyecto y en qué orden.

- *Característica 4.4. ¿Se realiza una estimación de la duración de cada actividad?*

  Consiste en estimar la duración de cada una de las actividades específicas que se hayan definido. Para este cálculo deberán tenerse en cuenta los recursos humanos disponibles y las habilidades de los mismos.

  Resultado esperado: Se efectúa una estimación de la duración de cada actividad.

- *Característica 4.5. ¿Se construye un cronograma para el proyecto?*

  En un cronograma se establece una fecha de inicio y fin para cada una de las actividades específicas, teniendo en cuenta su duración estimada. Este cronograma debe ser lo más realista posible, para evitar retrasos en las entregas del proyecto.

Resultado esperado: Se construye un cronograma donde se estima la fecha de inicio y fin de cada actividad.

- *Característica 4.6. ¿Existen actividades de control del cronograma?*

  Este aspecto se refiere a la revisión periódica del estado actual del proyecto para determinar si es necesario modificar y actualizar el cronograma del mismo.

  Resultado esperado: Se definen actividades de revisión del cronograma a lo largo del proyecto.

Gestión del costo

Las actividades en esta área se encargan de la planificación y control de los costos del proyecto. La planificación se materializa mediante la estimación de costos y la creación de un presupuesto, el cual debe ser controlado para no excederse sobre los valores aprobados.

- *Característica 4.7. ¿Se efectúa una estimación de los recursos afectados por cada actividad?*

  Consiste en determinar qué recursos (materiales, herramientas, recursos humanos, etc) y en qué cantidad se necesitan para llevar a cabo cada una de las actividades del proyecto. Esta acción posibilita la posterior estimación de costos.

  Resultado esperado: Se estiman los recursos que requiere cada actividad del proyecto.

- *Característica 4.8. ¿Se realiza una estimación de los costos del proyecto?*

  La estimación de los costos del proyecto se efectúa calculando el costo de los recursos necesarios para cada actividad. Para esta estimación es importante suponer diferentes escenarios, ya que algunos costos probablemente cambien durante el transcurso del proyecto.

  Resultado esperado: Se estiman los costos de los recursos requeridos por cada actividad.

- *Característica 4.9. ¿Se construye un presupuesto de costos?*

  El presupuesto consiste en la suma de los costos estimados de las actividades del cronograma, para establecer una línea base de costo total y poder medir el rendimiento del proyecto [25].

  Resultado esperado: Se construye un presupuesto de costos para el proyecto.

- *Característica 4.10. ¿Existen actividades de control del presupuesto a medida que avanza el proyecto?*

  Estas actividades consisten en realizar un seguimiento de los costos actuales del proyecto para detectar desvíos respecto a la planificación del presupuesto. Una de las técnicas más utilizadas es la del Valor Ganado (Earned Value Analysis), que permite comparar el costo real contra el costo planeado para el trabajo terminado.

  Resultado esperado: Se proponen actividades de control del presupuesto del proyecto.

Gestión del equipo de trabajo

Esta área se refiere a todos aquellos procesos que organizan y dirigen a los recursos humanos del proyecto. El equipo de trabajo está formado por las personas a las que se le han asignado roles y responsabilidades en las actividades del cronograma.

- *Característica 4.11. ¿Se efectúa una planificación de los recursos humanos?*

  La planificación de los recursos humanos se refiere a la identificación y asignación de los roles y responsabilidades dentro del equipo de trabajo. Los roles pueden ser asignados a personas o grupos. Entre las técnicas de planificación más utilizadas se encuentran las matrices de asignación de responsabilidades, las cuales detallan para cada actividad qué miembros del equipo quedan afectados y en qué nivel de responsabilidad.

  Resultado esperado: Se planifica claramente qué responsabilidad y rol tendrá cada uno de los miembros del equipo de trabajo.

- *Característica 4.12. ¿Se proponen actividades para motivar la interacción entre los miembros del equipo?*

  El trabajo en equipo requiere que se mantenga una buena comunicación entre los integrantes del proyecto. Las reuniones periódicas, la resolución grupal de conflictos y la conversación son técnicas muy útiles para mantener la interacción constantemente.

  Resultado esperado: Se planifican reuniones periódicas entre los miembros del equipo de trabajo.

- *Característica 4.13. ¿Se efectúa un seguimiento del rendimiento de los recursos humanos?*

  El seguimiento del rendimiento de los recursos humanos aumenta las probabilidades de cumplir con los objetivos del proyecto. Al evaluar

periódicamente (formal o informalmente) el rendimiento del equipo se descubre si es necesario establecer nuevas políticas de comunicación, motivación o reconocimiento.

Resultado esperado: Se evalúa periódicamente el rendimiento de los recursos humanos asignados al proyecto.

Dirección del riesgo

Un riesgo en un proyecto es un evento o condición incierta que, si se produce, tiene un efecto positivo o negativo sobre al menos un objetivo del proyecto, como tiempo, costo, alcance o calidad y se mide en términos de sus consecuencias y probabilidad [25]. La gestión del riesgo incluye la identificación de los potenciales riesgos para el proyecto, la evaluación de sus impactos, sus probabilidades de ocurrencia, y finalmente la priorización de los mismos. El análisis de riesgos es una actividad que debe ser continua durante el transcurso del proyecto, evaluando el surgimiento de nuevos riesgos o bien la modificación de los ya identificados.

- *Característica 4.14. ¿Se efectúa una identificación de los riesgos del proyecto?*

  Consiste en identificar que riesgos podrían afectar al proyecto y documentar las características de cada uno. En el proceso de identificación de riesgos deberían considerarse riesgos internos y externos. Internos son aquellos que pueden ser controlados e influenciados por el equipo del proyecto, como por ejemplo la estimación del costo. Riesgos externos son aquellas cuestiones que quedan fuera del alcance del equipo de proyecto, como cambios en el entorno.

  Resultado esperado: Se identifican los potenciales riesgos, internos y externos, que pueden afectar al proyecto.

- *Característica 4.15. ¿Se realiza una cuantificación y priorización de los riesgos?*

  A cada riesgo se le asigna una probabilidad de ocurrencia y se estima su impacto. Se debe tener en cuenta la interacción que pudiera existir entre diferentes riesgos. Esta jerarquización resulta de gran importancia, ya que durante el desarrollo del proyecto será importante focalizar la atención sobre la gestión de los riesgos con mayor prioridad.

  Resultado esperado: Se construye una lista priorizada de riesgos cuantificados.

- *Característica 4.16. ¿Se planifican acciones de respuesta ante cada riesgo?*

Consiste en determinar estrategias de prevención del riesgo (medidas proactivas) y de respuesta (medidas reactivas). Se deben definir los responsables de dichas actividades.

Resultado esperado: Se definen medidas proactivas y reactivas para cada riesgo.

- *Característica 4.17. ¿Existen actividades de supervisión y control de los riesgos?*

  A medida que transcurre el proyecto el listado de riesgos cuantificados puede ir cambiando, motivo por el cual resulta de gran importancia mantenerlo actualizado (identificando nuevos riesgos y modificando aquellos que hayan cambiado su impacto).

  Resultado esperado: Se proponen actividades de supervisión de los riesgos del proyecto.

3.5. Consideraciones sobre la utilización del marco comparativo

En este capítulo se ha desarrollado un marco comparativo el cual propone la evaluación de un conjunto de características para llegar a confrontar metodologías de minería de datos.

Como se ejemplifica en la Tabla 8, las características podrán ser evaluadas positiva o negativamente dependiendo si las metodologías en estudio cumplen o no con las mismas, obteniendo como resultado final el porcentaje de valoraciones positivas de cada enfoque.

No se ha realizado una ponderación de cada ítem debido a la subjetividad que podría implicar dicha tarea, quedando esta decisión a criterio del usuario de este marco comparativo. Si se utilizara un sistema de puntajes los resultados probablemente sean diferentes (ya que la presencia de cada característica tendrá su propio peso).

Característica	**Metodología 1**	**Metodología 2**	**Metodología 3**
Nivel de detalle en las actividades de cada fase			
Característica 1.1	NO	NO	SI
Característica …	SI	NO	SI
Escenarios y puntos de partida del proyecto	..	..	..
…	..	..	..
Total de características cumplidas	40 de 52 (77%)	35 de 52 (67%)	51 de 52 (98%)

Tabla 8. Ejemplo de comparación entre 3 metodologías.

4. Un caso de estudio

En el capítulo 2, luego de realizar una descripción de KDD, SEMMA, CRISP-DM y Catalyst, se ha concluido que sólo los dos últimos enfoques deberían ser considerados una metodología de minería de datos, ya que además de describir las actividades específicas de cada fase proveen una guía de cómo llevar a cabo el trabajo.

En este capítulo se aplicarán las metodologías CRISP-DM y Catalyst al caso de un centro médico para caracterizar las mismas desde un punto de vista práctico, dejando en segundo plano el estudio de las técnicas y algoritmos que se utilizan durante el proceso. Los resultados que se muestran a continuación, sintetizan y resumen la realización de las diferentes tareas en ambas metodologías.

4.1. Descripción del caso de estudio

En la actualidad, en los centros de salud existen ciertas especialidades médicas con agenda de turnos saturada. En algunos casos, cuando un paciente necesita ser atendido, no logra conseguir un turno (atención programada) para una fecha a corto plazo, debiendo esperar varios días para poder asistir a una consulta médica.

Esta situación se genera, fundamentalmente, debido al reducido horario de atención de algunos médicos y a la gran demanda de atenciones médicas especializadas. El problema crece cuando existe una alta tasa de ausentismo, ya sea porque el paciente no asiste a la consulta, o bien porque cancela su turno sin anticipación, el mismo día de la atención.

El caso en estudio es el de un centro médico ubicado en la provincia de Santa Fe, Argentina. En el mismo atienden seis médicos, todos de la misma especialidad (Clínica).

Los directivos del centro médico están interesados en reducir la tasa de ausentismo para aprovechar mejor los horarios de atención de los profesionales.

Las metodologías se han aplicado al caso de estudio tomando como eje las etapas generales de un proyecto de minería de datos (análisis del negocio, selección y preparación de los datos, modelado, evaluación, implementación). Para cada fase, se han ejecutado primero las actividades propuestas por CRISP-DM y luego las propuestas por Catalyst (Tabla 9).

Fases	CRISP – DM	CATALYST
Análisis y comprensión del negocio	Comprensión del negocio	Modelado del negocio
Selección y preparación de los datos	Entendimiento de los datos	Preparación de los datos
	Preparación de los datos	
Modelado	Modelado	Selección de herramientas y modelado inicial
Evaluación	Evaluación	Refinamiento del modelo
Implementación	Despliegue	Comunicación

Tabla 9. Paralelismo entre las fases de CRISP-DM y Catalyst.

4.2. Análisis y Comprensión del Negocio

4.2.1. Análisis y Comprensión del Negocio con CRISP-DM

En CRISP-DM, este análisis se lleva a cabo en la primera fase de la metodología llamada "Comprensión del negocio", la cual propone la realización de cuatro tareas: determinar objetivos del negocio, evaluar la situación, determinar objetivos de la minería de datos y crear un plan del proyecto. El resultado de cada tarea se representa a través de las salidas o entregables.

Tarea 1. Determinar los objetivos del negocio

Background

El centro médico en estudio recibe diariamente una gran cantidad de pacientes, los cuales asisten con un turno previo. En el centro trabajan seis médicos clínicos de lunes a viernes, desde las 8:00 hs. hasta las 19:00 hs.

La Figura 12 representa la estructura de la organización.

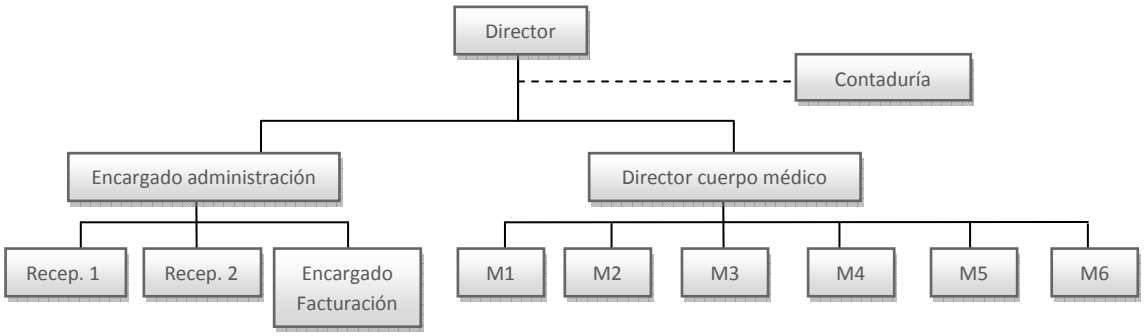

Fig. 12. Organigrama del centro médico.

Entre las personas clave de la organización encontramos:

- Director del centro médico: persona encargada de la coordinación, administración y gestión integral de la organización.
- Encargado de administración: persona que supervisa y coordina a los recepcionistas y al encargado de facturación.
- Director del cuerpo médico.
- Recepcionistas: personal que otorga y recepciona los turnos de los pacientes.
- Encargado de facturación.
- Médicos (M1, M2, M3, M4, M5, M6).

El director del centro estima que aproximadamente el 20% de los pacientes que solicitan turno no asisten a su consulta, lo que provoca que otros pacientes no puedan aprovechar el mismo.

Los integrantes del nivel gerencial (director, encargado de administración y encargado del cuerpo médico) tienen poco conocimiento en materia de análisis de datos, pero tienen presente que la minería de datos puede colaborar a la solución de su problema. Sus expectativas se centran en mejorar el servicio al paciente, pudiéndole ofrecer una fecha de atención lo más pronta posible con los recursos médicos disponibles. Una forma de llevar a cabo este objetivo es detectando que pacientes no asistirán a su turno, para otorgar el mismo a otra persona.

Actualmente la solución que se implementa es el otorgamiento de sobreturnos, para compensar aquellos pacientes que pudieran llegar a faltar. Esta solución por lo general no es efectiva, ya que la cantidad de sobreturnos que se otorgan depende del criterio del recepcionista, provocando la saturación de pacientes en el consultorio. Si bien como ventaja genera la atención temprana del paciente, provoca la queja de los médicos porque deben quedarse más tiempo, fuera de su horario laboral, y trabajar con apuros.

Objetivos del negocio

El proyecto de minería de datos tiene como objetivo mejorar la gestión de turnos médicos, ya que brindará información acerca de qué pacientes probablemente no asistan a su turno. Con esta información se podrán tomar acciones preventivas, como confirmaciones telefónicas con aquellos pacientes con alta probabilidad de ausentarse.

Desde una perspectiva de negocio se espera:

- Reducir el nivel de pacientes que no asisten a su consulta, mejorando el aprovechamiento del horario de atención del médico.
- Ofrecer turnos a los pacientes lo antes posible, sin sobrecargar la agenda del médico.

Criterios de éxito

Se considera como criterio de éxito para el proyecto la reducción del ausentismo de pacientes en un 50%.

Este criterio es evaluado por el sector gerencial de la organización (directores y encargados).

Tarea 2. Evaluación de la situación

Inventario de recursos

Recursos de hardware	• Dos PCs y una notebook, todas con procesador doble núcleo y 2 GB de RAM. • Impresora Láser.
Recursos de Software	Se trabajará con herramientas de Software libre • WEKA y R para el modelado de los datos. • OpenOffice para la generación de los reportes.
Recursos de datos y conocimiento	• Se utilizará la base de datos del sistema operacional de la organización (sistema de turnos), la cual corre sobre la plataforma PostgreSQL.
Recursos humanos	• 1 analista de explotación de información.

Tabla 10. Inventario de recursos

Requerimientos, supuestos y restricciones

Requerimientos

- Mantener la confidencialidad de los datos de pacientes y médicos.
- Los resultados del proceso deben estar representados de una forma clara, simple y entendible para los usuarios.

Supuestos

- Los turnos registrados en la base de datos son reales.
- Las fuentes de datos están libres de errores y son accesibles en todo momento.

- Se contará con la colaboración del personal gerencial en todo momento.

Restricciones

- El proyecto deberá limitarse a la utilización de los recursos citados en el inventario, y no requerir insumos extra.

Gestión del riesgo

El impacto de un riesgo puede ser alto / medio / bajo[5].

Prioridad	Descripción del riesgo	Probabilidad	Impacto
1	Los patrones encontrados no logran satisfacer los objetivos del proyecto.	0.5	Alto
2	El análisis exploratorio de los datos indica baja calidad de los mismos, dificultando la aplicación de técnicas de minería.	0.5	Alto
3	Los patrones encontrados no son entendibles por la gerencia.	0.4	Medio
4	La gerencia no participa activamente en el proyecto.	0.3	Medio

Tabla 11. Lista de riesgos del proyecto

Planes de contingencia

Riesgo	**Acciones de contingencia**	
	Medidas proactivas	**Medidas reactivas**
Los patrones encontrados no logran satisfacer los objetivos del proyecto.	• Utilizar distintas técnicas de minería en la fase de modelado. • Utilizar distintos parámetros para los modelos obtenidos.	• Recolectar mayor cantidad de datos.
El análisis exploratorio de los datos indica baja calidad de los mismos dificultando la aplicación de técnicas de minería.	• Recolectar datos que representen lo mejor posible la realidad del problema.	• Buscar otras fuentes de datos. • Recolectar mayor cantidad de datos (variables u observaciones según el caso).
Los patrones encontrados no son entendibles por la gerencia.	• Capacitar al personal en análisis de datos y de resultados.	• Armar nuevas representaciones de los patrones encontrados.
La gerencia no participa activamente en el proyecto.	• Señalar a la gerencia la importancia de su compromiso desde etapas tempranas del proyecto. • Informar permanentemente a la gerencia acerca de los avances del proyecto. • Escuchar todas sus sugerencias para que se sientan parte del mismo.	• Determinar las causas de la falta de interés y trabajar sobre las mismas.

Tabla 12. Plan de contingencia

[5] CRISP-DM sólo propone la identificación de los riesgos. En este trabajo se ha estimado adicionalmente la probabilidad de ocurrencia y el impacto de cada uno.

Glosarios

Terminología del negocio:

- Turno (o atención programada): es una reserva que un paciente realiza para ser atendido en un día y horario específico, por un profesional del centro médico.
- Ausentismo: representa la inasistencia del paciente a su turno.

Terminología de minería de datos:

- Métodos/técnicas de minería de datos: Son herramientas utilizadas para buscar los patrones. Un método puede ser descriptivo o predictivo.
    - Métodos descriptivos: Proporcionan información sobre las relaciones existentes entre los datos. Exploran las propiedades de los datos.
    - Métodos predictivos: responden a preguntas sobre datos futuros. Permiten estimar el valor futuro de variables "dependientes" (o explicadas), a partir de otras llamadas "independientes" (o explicativas).
- Vista minable: estructura final de los datos, con formato tabular, que se proporciona como entrada a los algoritmos de minería. Las filas representan las observaciones y las columnas las variables en estudio.
- Datos de entrenamiento/prueba: los datos de entrenamiento son aquellos que se utilizan para armar el modelo de minería, mientras que los de prueba se utilizan para testear la calidad del mismo.

Análisis de costo/beneficio

En este caso, como el proyecto de minería se desarrolla en el contexto de un trabajo de investigación, no se realiza un análisis costo/beneficio.

Tarea 3. Determinar los objetivos de la minería de datos

Objetivo de la minería de datos

Se trabajará con algoritmos de clasificación (como árboles o regresión logística) para construir un modelo predictivo, que permita estimar si un paciente asistirá o no a la consulta.

Criterio de éxito para el proyecto de minería

Se espera que la capacidad predictiva del modelo sea buena, con una tasa de acierto de al menos el 70%.

Tarea 4. Crear el plan para el proyecto de minería de datos

Plan del proyecto

La duración estimada para la ejecución el proyecto es de 18 semanas. Se espera que la fase de evaluación del modelo sea exitosa, y no requiera retroceder a etapas anteriores. El único recurso humano que trabajará en el proyecto será un analista, 15 hs semanales.

En la Tabla 13 se representa el cronograma del proyecto. Se puede observar que contempla una interacción entre la fase de preparación de los datos y el modelado.

Evaluación inicial de técnicas y herramientas

El proyecto se llevará a cabo con herramientas de software libre. Para la fase de análisis de datos y modelado se utilizarán los paquetes de software R [28] y Weka [34].

Para la selección de las técnicas de clasificación que se utilizarán, resulta importante la naturaleza de los datos. Como en nuestro análisis las variables explicativas son mixtas (numéricas y cualitativas) y la explicada es binaria (presente/ausente), las técnicas candidatas son:

- Árboles de decisión
- Regresión logística
- Clasificador Naive Bayes
- Método de vecino más próximo (KNN)

Actividad	Semana																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Entendimiento del negocio																		
1.1. Determinar los objetivos del negocio																		
1.2. Evaluar la situación																		
1.3. Determinar los objetivos de la minería de datos																		
1.4. Crear un plan para el proyecto de minería de datos.																		
2. Comprensión de los datos																		
2.1. Recolectar los datos iniciales																		
2.2. Describir los datos																		
2.3. Explorar los datos																		
2.4. Verificar la calidad de los datos.																		
3. Preparación de los datos																		
3.1. Seleccionar los datos																		
3.2. Limpieza de datos																		
3.3 Construcción de los datos																		
3.4. Integrar los datos																		
3.5. Formatear los datos																		
4. Modelado																		
4.1 Seleccionar la técnica de modelado																		
4.2. Diseñar las pruebas del modelo																		
4.3. Construir el modelo																		
4.4. Evaluar el modelo																		
5. Evaluación																		
5.1. Evaluar los resultados																		
5.2. Revisión del proceso																		
5.3. Determinar las siguientes etapas																		
6. Implementación																		
6.1. Planificar la implementación																		
6.2. Planificar el monitoreo y el mantenimiento																		
6.3. Crear el reporte final. Revisión del proyecto																		

Tabla 13. Cronograma del proyecto.

4.2.2. Análisis y Comprensión del Negocio con Catalyst

La metodología Catalyst implementa tareas para el análisis y comprensión del negocio en su primera parte: Metodología para el Modelado del Negocio. El análisis comienza identificando el escenario que constituye el punto de partida para nuestro proyecto. En el caso de estudio, el escenario que se adecúa es aquel que tiene por objetivo colaborar en la solución de un problema organizacional. A continuación se llevan a cabo las tareas específicas planteadas para este escenario.

Tarea 1. Identificar y caracterizar al personal interesado (stackeholders)

Las personas involucradas en el proyecto según la clasificación propuesta por la metodología son:

- Personas que viven día a día el problema organizacional: médicos, recepcionistas.
- Personas que financian el proyecto: director.
- Personas que toman decisiones acerca de la continuidad del proyecto: director.
- Personas que determinan si el proyecto es exitoso: encargado de administración.
- Personas que se beneficiarán con los resultados del proyecto: recepcionistas y médicos.

Tarea 2. Entrevistar y explorar la situación de negocio con el personal interesado

Entre las preguntas más importantes que se realizaron al personal involucrado encontramos:

- Recepcionistas
  - ¿Recibe quejas de los pacientes cuando solicitan un turno? ¿Cuáles?
  - ¿Observa un alto nivel de ausentismo en las consultas?
  - ¿Existen momentos donde el profesional no esté atendiendo porque faltaron pacientes?
  - ¿Cómo reduciría la cantidad de pacientes que no asisten a su turno?
  - ¿Le sería útil alguna herramienta que le permita identificar que pacientes probablemente no concurran a su turno?
- Director y encargados
  - ¿Cómo definirían la problemática actual con los turnos?
  - ¿Qué soluciones ya han evaluado? ¿Cuáles fueron los resultados?

- ¿Qué expectativas tienen para el proyecto? ¿Cuál sería un buen resultado?

- Médicos
  - ¿Están conformes con la organización actual de los turnos? ¿Qué mejoras introduciría?
  - ¿Recibe quejas de los pacientes con respecto a la gestión de los turnos? ¿Cuáles?

Las respuestas a estas preguntas se han redactado en una narrativa que se detalla en la tarea siguiente.

Tarea 3. Enmarcar la situación del negocio

Actualmente, los recepcionistas y la dirección coinciden en que si bien la cantidad de pacientes ausentes por día no es muy grande (empíricamente se sabe que se encuentra entre el 15% y 20%), es suficiente como para que otros pacientes puedan aprovechar esos turnos.

Los directivos del centro médico sostienen que es importante ofrecerle al paciente una fecha de atención lo más pronta posible (independientemente de si la toma o no). Con frecuencia el personal de recepción recibe quejas de los pacientes por la cantidad de días que deben esperar hasta ser atendidos.

Los médicos han manifestado que están conformes con la organización actual de los turnos, aunque en ocasiones el personal de recepción otorga muchos sobreturnos, causando la saturación de pacientes en el consultorio.

El personal interesado espera contar con información acerca de qué pacientes probablemente no asistan a su turno, para tomar medidas proactivas como confirmaciones telefónicas o bien sobreturnos adicionales basados en un criterio objetivo y probabilístico. Se espera que el modelo de minería colabore en esta tarea, brindando una predicción sobre la asistencia del paciente, con una determinada tasa de acierto.

Tarea 4. Identificar los objetivos de negocio relevantes para el proyecto.

En la Figura 13, se detallan los objetivos de negocio relevantes para el proyecto y la relación que existe entre ellos.

El principal objetivo es brindar una buena atención a los pacientes. Este objetivo se concreta a partir de otros objetivos como atender correctamente al paciente durante la consulta y ofrecer una fecha de atención lo más pronta posible. Con éste último colabora el proyecto de minería de datos.

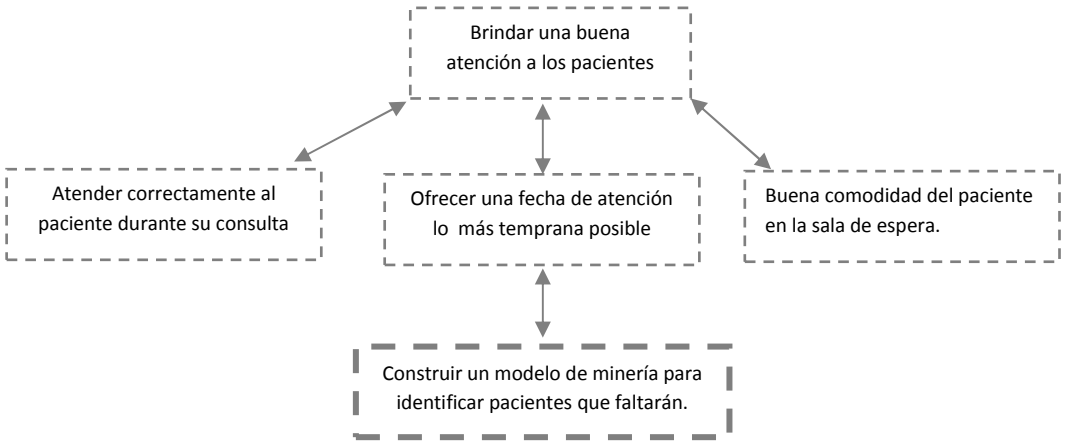

Fig. 13. Objetivos del negocio relevantes para el proyecto.

Tarea 5. Buscar los datos que se explorarán

Los datos serán extraídos de la base de datos del sistema operacional de turnos. El periodo en estudio será desde el 01/01/2011 al 31/07/2011.

Las unidades observacionales del problema son aquellos turnos otorgados que no fueron cancelados.

No se requerirá en esta etapa la integración de distintas fuentes, ya que toda la información reside en esta base de datos.

La estructura de la base de datos contiene 68 tablas. Sólo 2 de ellas (tabla **turno** y tabla **paciente**) contienen información relevante para el problema en estudio (Fig. 14).

En cuanto al proceso de solicitud de turno es el siguiente:

Camino básico

1. El paciente solicita al recepcionista un turno para un médico.
2. El recepcionista propone una fecha y horario de atención.
3. El paciente acepta el horario propuesto.

Cuando llega el día de la atención

4. El paciente informa su llegada en la recepción.
5. La recepcionista registra la llegada del paciente y le informa al médico.
6. El médico llama al paciente para su atención.

Caminos alternativos

3. El paciente no acepta la fecha/hora propuesta y solicita otro horario de atención

    3.1. La recepcionista propone un nuevo horario de atención

4. El paciente no asiste a la consulta. El turno queda sin utilizarse.

El proyecto de minería de datos apunta a mejorar el camino alternativo del cuarto paso en el proceso (el paciente no asiste a la consulta).

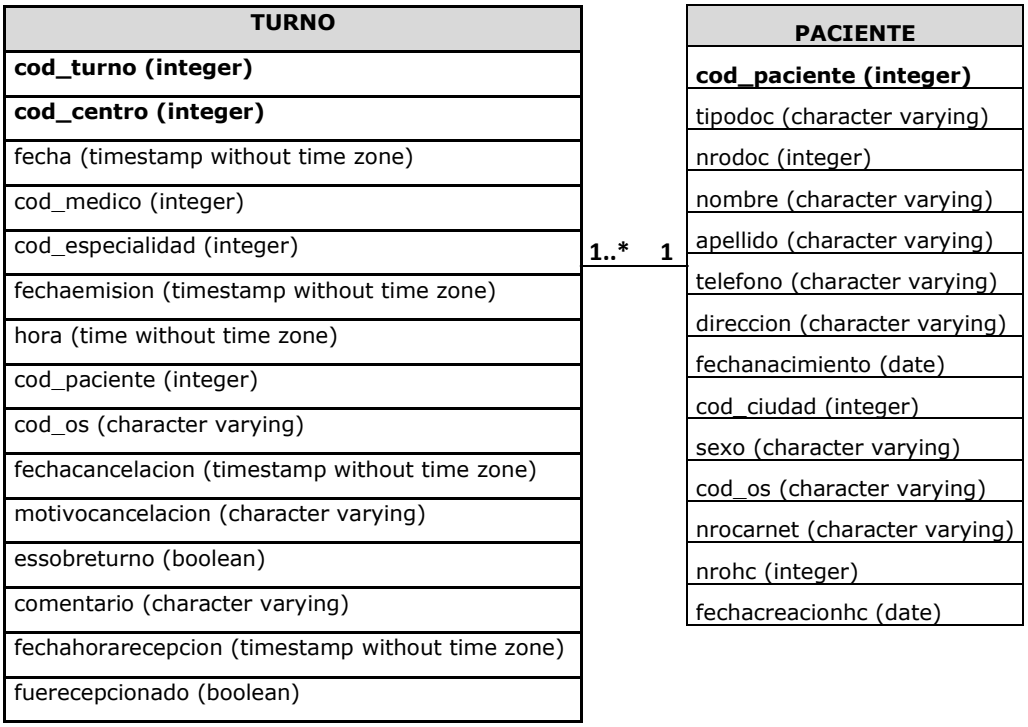

Fig. 14. Estructura de las tablas Turno y Paciente

Tarea 6. Armar el caso del negocio

Este documento representa un plan de proyecto.

Problemática actual de la organización.

Actualmente el centro médico atiende una gran cantidad de pacientes, muchos de los cuales no consiguen un turno a corto plazo. Para dar solución a esta situación, en ocasiones los recepcionistas otorgan sobreturnos basándose en su experiencia, situación que muchas veces provoca la saturación de la agenda del médico. A su vez, muchos de los pacientes que poseen turno no asisten a su consulta, provocando que otras personas no puedan aprovechar el mismo.

Las alternativas de solución al problema.

1. Implementar un proyecto de minería de datos para estimar potenciales pacientes ausentes. Con un modelo de minería de datos se estimarán que pacientes probablemente no asistan a su consulta, lo que permitirá tomar acciones preventivas (como confirmar su atención telefónicamente o dar un sobreturno en su lugar).
2. Contratar más médicos. Luego de un estudio económico y financiero los directivos han descartado esta alternativa.
3. Llamar telefónicamente a todos los pacientes para confirmar su cita. Esta alternativa resulta poco viable, ya que además del costo de las

llamadas telefónicas, se ocupa tiempo de los recepcionistas que intentan contactar al paciente.

4. No hacer nada. Con esta alternativa la situación del negocio no mejorará, provocando la insatisfacción de aquellos pacientes que requieren una pronta atención (los cuales en ocasiones asistirán a otro centro médico).

La solución propuesta.

Un modelo de minería de datos que permita estimar los pacientes que no asistirán a su consulta. El modelo será implementado dentro del sistema operacional de turnos, notificándoles a los recepcionistas de esta probable situación. Con esta información se podrá confirmar previamente la cita con el paciente, o bien dar un sobreturno en su lugar para cubrir el mismo.

Recursos que se utilizarán. Cronograma de tiempos.

Este punto ya se ha evaluado en la tarea 4 de la fase "Comprensión del Negocio" en CRISP-DM.

Análisis financiero ROI (retorno de la inversión).

Este punto ya se ha evaluado en la tarea 2 de la fase "Comprensión del Negocio" en CRISP-DM.

Tarea 7. Presentar el caso de negocio al personal interesado

El caso de negocio fue presentado a todo el personal involucrado, el cual se mostró conforme con el proyecto.

Tarea 8. Describir la situación del negocio para el proceso de minería

Los usuarios están dispuestos a participar activamente del proyecto, ya que comprenden los beneficios del mismo. Consideran importante que el modelo sea implementado en el contexto del sistema operacional de turnos, para poder tener toda la información integrada. Gracias a un trabajo conjunto con los desarrolladores del software de turnos, esta acción es posible de llevar a cabo.

Análisis del riesgo.

Este punto ya se ha evaluado en la tarea 2 de la fase "Comprensión del Negocio" en CRISP-DM.

Tarea 9. Definir los requerimientos de la implementación

Entrega y distribución del modelo

El modelo resultante del proceso se implementará en el sistema de gestión de turnos de la clínica, tarea que se realizará conjuntamente con el desarrollador del sistema.

Identificación de potenciales usuarios del modelo

Personal de recepción y administración.

Capacitación a los potenciales usuarios del modelo

Se realizará un plan de capacitación breve para explicar las salidas del modelo y los potenciales usos de la misma.

Medición de los efectos del modelo

Una vez implementado el modelo se procederá a comparar la nueva tasa de ausentismo con los valores históricos de la misma.

Determinación de la caducidad del modelo (cuándo dejará de ser válido)

El modelo podría dejar de ser válido cuando:

- Se contraten más médicos.
- Factores externos a la organización provoquen un cambio en el hábito de asistencia de los pacientes a su turno.

En estos casos, se requerirá probar nuevamente el modelo bajo las nuevas condiciones para validar su vigencia.

Mantenimiento del modelo

El mantenimiento que requerirá el modelo será la realización de pruebas periódicas sobre la tasa de acierto, para verificar que se mantenga en los márgenes de tolerancia.

Documentación requerida

Para el personal de recepción se entregará un instructivo que explique la utilidad del modelo y cómo se implementa en el sistema de turnos.

Para la gerencia se creará un resumen del proyecto, donde se explicite los datos analizados, los resultados y la forma en la que los mismos fueron implementados en la operatoria diaria de la organización.

4.3. Selección y Preparación de los Datos

4.3.1. Selección y Preparación de los Datos con CRISP-DM

En CRISP-DM, estas actividades se llevan a cabo en dos fases, llamadas "Comprensión de los datos" y "Preparación de los datos".

Comprensión de los datos

Tarea 1. Recolectar los datos iniciales

Reporte inicial de recolección de datos

Este punto ya se ha efectuado en la tarea 5 de la etapa "Modelado del Negocio" en la metodología Catalyst.

Tarea 2. Describir los datos

Reporte de descripción de los datos

Para obtener la información de la base de datos, se ha ejecutado la consulta SQL de la Figura 15.

```
SELECT
fecha,
date_part('hour',hora) as hora,
to_char(fecha,'DAY') AS dia,
'm' || cod_medico as medico,
CASE WHEN fechaCreacionHC is null THEN false ELSE
CASE WHEN fecha < fechaCreacionHC THEN false ELSE true END END as tienehc,
esSobreturno,
date_part('year', age(date(turno.fecha),date(paciente.fechanacimiento))) as edad,
sexo,
CASE WHEN paciente.cod_ciudad=1 THEN true ELSE false END as esdelaciudad,
fechaemision,
CASE WHEN turno.cod_os is null THEN false ELSE true END as atencion_os,
CASE WHEN fechahorarecepcion is null THEN 'NO' ELSE 'SI' END as asistio
FROM
turno INNER JOIN paciente ON turno.cod_paciente=paciente.cod_paciente
WHERE fecha > '2011-01-01' and fecha < '2011-07-31' and fechacancelacion is null
```

Fig 15. Consulta SQL utilizada para extraer los datos

La consulta arrojó como resultado **8810** registros (turnos), de los cuales se obtuvo:

- Fecha del turno: fecha para la cual se le otorgó un turno al paciente.
- Hora: hora del turno.
- Día: día de la semana del turno.
- Médico: médico que asistirá al paciente.
- Tiene HC: indica si el paciente tiene historia clínica al momento que solicita el turno.
- Es sobreturno: indica si el turno fue otorgado excepcionalmente por no haber más turnos disponibles.
- Edad del paciente.
- Sexo paciente.

- Es de la ciudad: indica si el paciente vive en la ciudad donde se ubica el centro médico (el código de la ciudad es 1).
- Fecha de emisión: fecha en la que el turno fue otorgado.
- Atención por obra social: indica si el paciente se atenderá por obra social.
- Asistió: indica si el paciente asistió a la consulta.

Atributo	Tipo	Valores posibles
Fecha turno	Fecha	[01/01/2011 – 31/07/2011]
Hora	Entero	[8-20]
Día	Categórica	[Lunes – Viernes]
Médico	Categórica	M1, M2, M3, M4, M5, M6
Tiene HC	Booleana	Si/no
Es sobreturno	Booleana	Si/no
Edad	Entero	
Sexo	Categórica	M / F
Es de la ciudad	Booleana	Si/no
Fecha de emisión	Fecha	
Atención OS	Booleana	Si/no
Asistió	Booleana	Si/no

Tabla 14. Descripción de las variables extraídas.

Tarea 3. Exploración inicial de los datos

Reporte de exploración de los datos

Los datos recolectados se han explorado con los paquetes de software WEKA y R.

La Figura 16 muestra la salida de WEKA, donde se puede analizar la distribución de cada una de las variables en estudio. Las instancias en color rojo representan los turnos de pacientes ausentes.

Analizando la distribución de la variable de respuesta (asistió) se puede observar que del total de turnos otorgados, en 1226 de ellos el paciente no ha asistido a la consulta (14% ausentismo).

Sin embargo, esta tasa resulta menor si discriminamos por las variables "es sobreturno" y "es de la ciudad". Para los pacientes que solicitan sobreturno la tasa de ausentismo es de 9.3% (Fig. 17). Para los pacientes que son de otra ciudad, la tasa de ausentismo se reduce al 6.3% (Fig. 18).

La tasa de ausentismo se reduce considerablemente para aquellos sobreturnos de pacientes que no son de la ciudad donde reside el centro médico (Fig. 19).

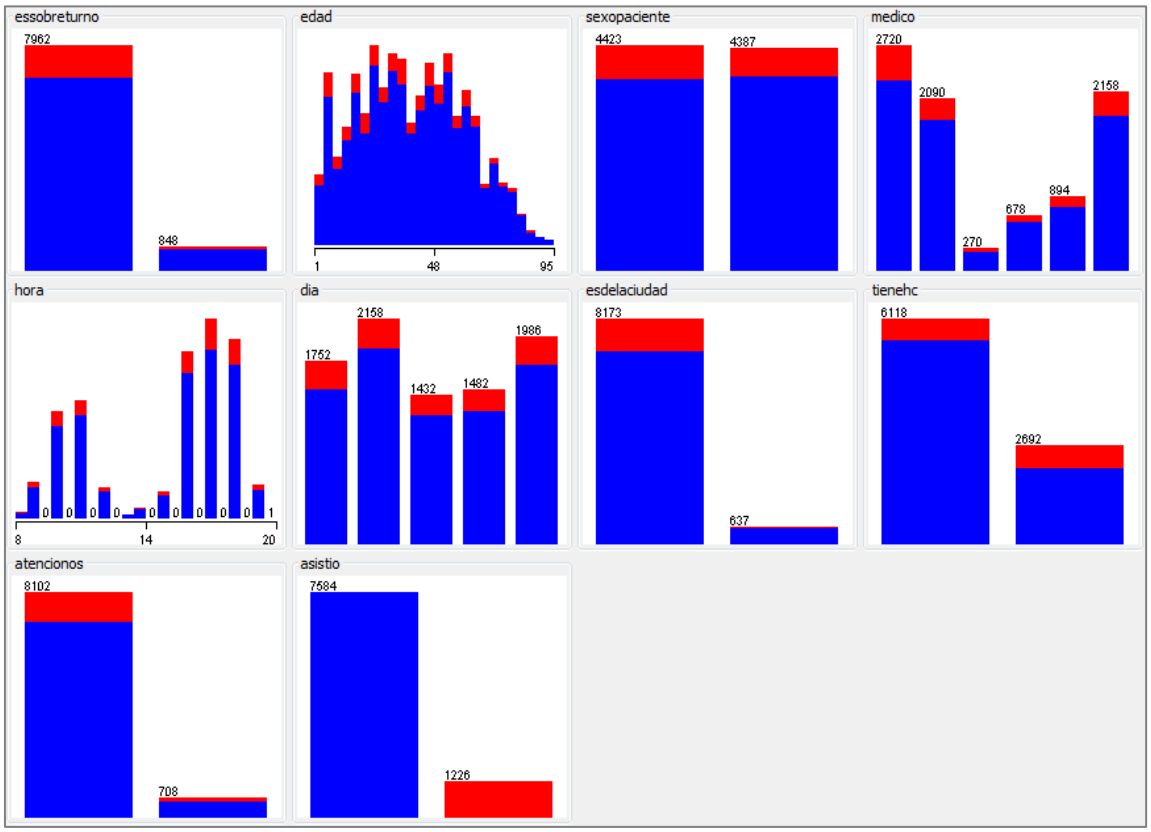

Fig. 16. Distribución de las variables del conjunto de datos.

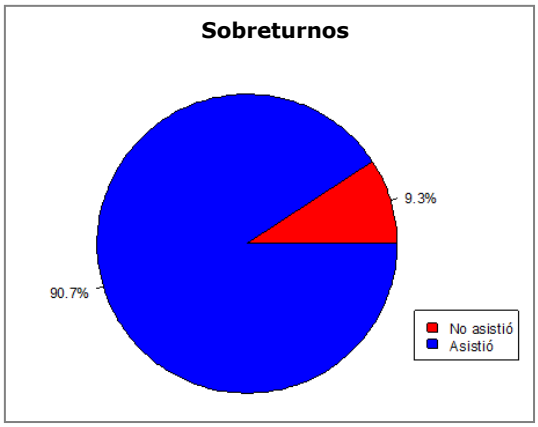

Fig.17. Tasa de ausentismo para los sobreturnos

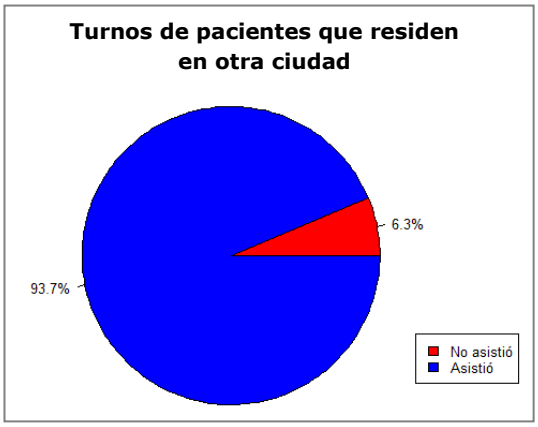

Fig 18. Tasa de ausentismo para pacientes de otra ciudad.

Fig. 19. Tasa de ausentismo para sobreturnos de pacientes que no son de la ciudad.

Es decir, que las variables "es de la ciudad" y "es sobreturno" tienen una alta relación con la asistencia del paciente. La variable "edad" tiene una media de 39 años y una mediana de 38 años (Fig. 20). Tiene 328 valores ausentes (4%).

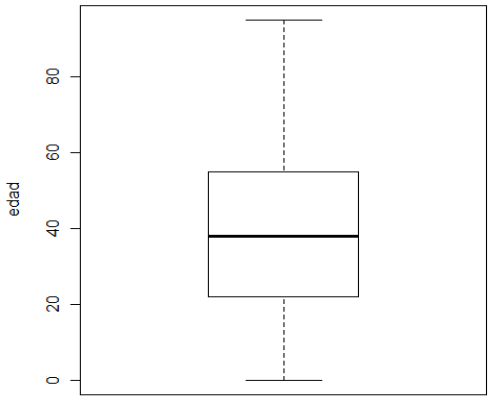

Fig. 20. Distribución de la variable edad.

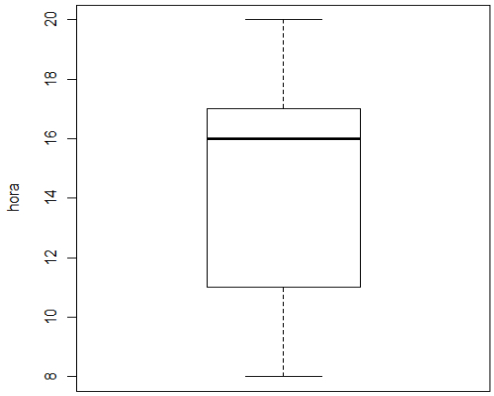

Fig. 21. Distribución de la variable hora.

El coeficiente de curtosis[6] para esta variable (obtenido con el software R) es de -0.8413558, por lo que no hay fuertes indicios de mezcla de poblaciones.

La variable "hora" tiene una media de casi 15 hs y una mediana de 16 hs (Fig. 21). Su coeficiente de curtosis resultó -1.241429, valor que confirma una mezcla de poblaciones, probablemente por un lado los turnos de la mañana (hasta las 13hs) y por otro los de la tarde (luego de las 13hs). En el histograma de la Figura 22 se observa claramente esta situación.

[6] El coeficiente de curtosis es un estadístico que estudia la distribución de frecuencias de una variable en su zona central. Es utilizado para analizar la homogeneidad de un conjunto de datos y detectar mezcla de poblaciones.

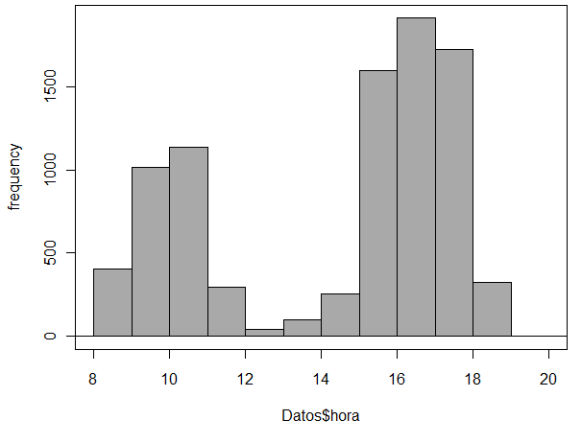

Fig. 22. Histograma de la variable hora.

La variable "fecha de emisión" no se evaluará en esta etapa, ya que más adelante será transformada para dar mejor soporte al problema en estudio.

Tarea 4. Verificar la calidad de los datos

Reporte de calidad de los datos

Los datos analizados no presentan grandes problemas en cuanto a su calidad.

La variable edad tiene 328 datos ausentes, donde la recepcionista no ha registrado la fecha de nacimiento del paciente.

En cuanto a los datos anómalos, las variables numéricas (hora y edad) no presentan datos extremos.

La variable de respuesta, "asistió", está formada por clases no balanceadas (ya que el 14% de los turnos resultaron ausentes). Este problema se abordará en la fase de modelado, utilizando matrices de costos para mejorar la performance de los algoritmos, y de ser necesario se utilizarán técnicas de muestreo para balancear las instancias.

Preparación de los datos

Tarea 1. Seleccionar los datos

No se descartará ninguno de los atributos obtenidos en la fase anterior.

Debido a que el clima es un factor muy importante que podría influir en el ausentismo de un paciente, se ha recolectado información acerca de las condiciones climáticas de cada día de atención. Mediante un sitio web que proporciona información climática histórica [33] se ha construido una nueva tabla que indica para cada día de atención si hubo precipitaciones a la mañana y a la tarde. Los campos de esta tabla son "Fecha", "Llovió por la mañana", "Llovió por la tarde". Estos datos se utilizarán para crear una nueva variable que indique si hubo precipitaciones para cada turno de la muestra.

Tarea 2. Limpieza de los datos

Reporte de limpieza de datos

En el reporte de calidad de los datos se señaló que 328 observaciones (aproximadamente el 4%) tienen ausente el valor de la edad. Se realizará una imputación de la mediana (38 años), ya que es una medida de posición más robusta que la media y menos sensible a los valores atípicos.

Tarea 3. Construcción de los datos

La variable "fecha de emisión del turno" será analizada junto con la fecha del turno, para obtener así los "días de antelación en la solicitud del turno".

Días antelación [días] = fecha turno – fecha emisión

La distribución de esta nueva variable está representada en la Figura 23 y en la Figura 24. Como se puede observar, la mayor parte de los turnos fueron otorgados a lo sumo quince días antes de la fecha de atención. Valores superiores a 16 podrían ser considerados anómalos.

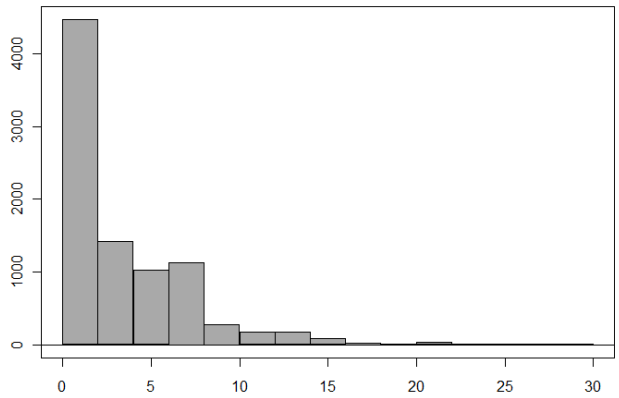

Fig. 23. Histograma de la variable "días antelación solicitud"

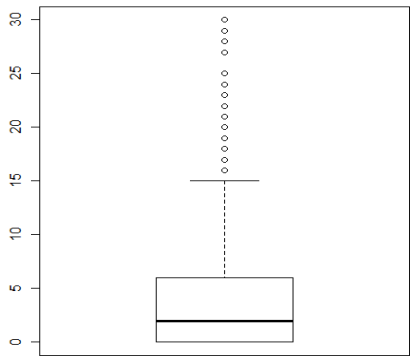

Fig. 24. Diagrama de cajas de la variable "días antelación solicitud"

Esta nueva variable calculada se encuentra relacionada con "es sobreturno", ya que como se puede apreciar en la Figura 25, la mayor parte de los sobreturnos han sido otorgados con menos de 6 días de antelación.

Por otro lado, a partir de la variable numérica "hora" se creará la variable categórica "horario", que asumirá los valores "mañana" (cuando la hora sea anterior a las 13hs) y "tarde"(cuando la hora sea posterior a las 13hs). La Figura 26 muestra la distribución dicha variable. La primer columna corresponde a los horarios de la tarde (5950 turnos) y la segunda a los de la mañana (2860 turnos). No se han observado relaciones significativas entre esta variable y las demás.

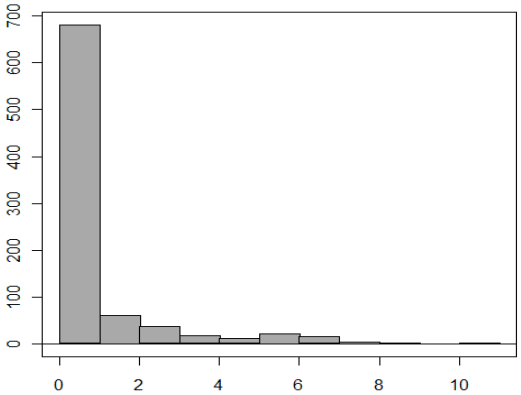

Fig. 25. Distribución de la variable "dias antelación solicitud" en los sobreturnos.

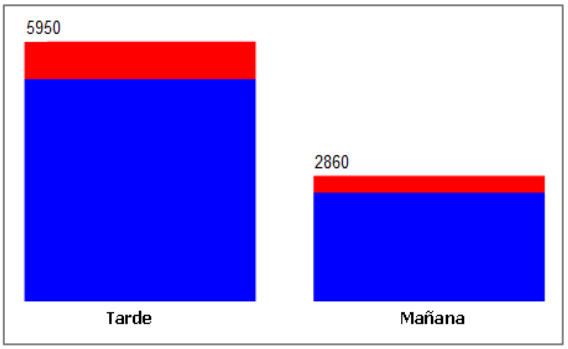

Fig. 26. Distribución de la nueva variable "horario".

Tarea 4. Integración de los datos.

Los datos de los turnos recolectados inicialmente se han integrado con la información externa de las precipitaciones. El resultado es un nuevo atributo llamado "llovió" que toma los valores "si" o "no". Se estima que en el 6% de los turnos registrados ha llovido (543 instancias).

Analizando la variable en función del ausentismo, se ha detectado que en los días en que llovió (Fig.27) la tasa de ausentismo fue de 17%, mientras que si no llovió (Fig.28) la tasa es del 14% (3% de diferencia).

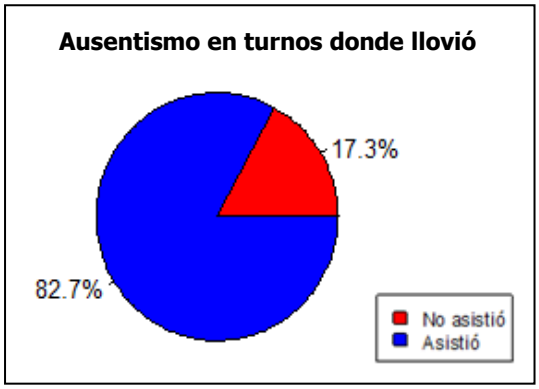

Fig. 27. Tasa de ausentismo cuando llovió.

Fig. 28. Tasa de ausentismo cuando no llovió.

Conjunto de datos final

El conjunto de datos final completo está formado por 8810 observaciones (turnos) caracterizadas por las variables descriptas en la Tabla 15. La última variable, "asistió", representa la salida del modelo (variable explicada).

Atributo	Tipo	Valores posibles
Horario	Categórica	[mañana-tarde]
Día	Categórica	[Lunes – Viernes]
Médico	Categórica	M1, M2, M3, M4, M5, M6
Tiene HC	Booleana	Si/no
Es sobreturno	Booleana	Si/no
Edad	Entero	
Sexo	Categórica	M / F
Es de la ciudad	Booleana	Si/no
Días antelación	Entero	[0-30]
Atención OS	Booleana	Si/no
Llovió	Booleana	Si/no
Asistió	**Booleana**	**Si/no**

Tabla 15. Variables del conjunto de datos final.

Resumen del conjunto de datos completo

En la Figura 29 se presenta un resumen del conjunto de datos final. La variable de respuesta ("asistió") no está balanceada, ya que la proporción de turnos a los que el paciente no asistió es mucho menor.

```
essobreturno      edad        sexopaciente medico      horario            dia
f:7962       Min.   : 1.00   F:4423       m1: 894   mañana:2860   FRIDAY   :1432
t: 848       1st Qu.:23.00   M:4387       m2:2158   tarde :5950   MONDAY   :1752
             Median :38.00                m3:2720                 THURSDAY :1986
             Mean   :39.23                m4: 678                 TUESDAY  :2158
             3rd Qu.:55.00                m5:2090                 WEDNESDAY:1482
             Max.   :95.00                m6: 270


 esdelaciudad tienehc  atencionos llovio   diasantelacionsolicitud asistio
 f: 637       f:2692   f: 708     f:8267   Min.   : 0.000          NO:1226
 t:8173       t:6118   t:8102     t: 543   1st Qu.: 0.000          SI:7584
                                          Median : 2.000
                                          Mean   : 3.577
                                          3rd Qu.: 6.000
                                          Max.   :30.000
```

Fig. 29. Resumen del conjunto de datos final.

4.3.2. Selección y Preparación de los Datos con Catalyst

En Catalyst, las actividades de selección y preparación de los datos se encuentran en la fase de "Preparación de los datos". La recolección inicial de los datos ya se ha realizado en la fase anterior (Modelado del Negocio).

Tarea 1. Caracterización de las variables

La caracterización de las variables consiste en hacer un análisis exploratorio de los atributos recolectados. Esta acción se ha realizado en las tareas 2 y 3 de la fase "Entendimiento de los datos" en CRISP-DM.

Tarea 2. Chequear problemas básicos en las variables

- *La variable tiene un valor único en todas las instancias.*

  No sucede.

- *La variable tiene el 80% o más de las instancias con datos ausentes.*

  No sucede.

- *La variable aparece numérica pero en realidad es una representación numérica de las categorías.*

  No sucede.

- *La variable tiene muchas categorías únicas (cientos o más).*
  No sucede.

Tarea 3. Chequear problemas básicos en el conjunto de datos.

No se han encontrado patrones inesperados durante el análisis exploratorio de los datos (realizado en la tarea 3 de la fase "Comprensión de los datos" en CRISP-DM).

Tarea 4. Chequear variables anacrónicas (que no aportan valor)

Luego del análisis exploratorio se concluye que no existe fuerte evidencia de que alguna de las variables no aporte información al modelo.

Tarea 5. Chequear que haya suficientes datos

Tras dividir el conjunto de datos en diferentes partes, se ha observado que el comportamiento de las variables es el mismo, por lo cual los datos son suficientes.

Tarea 6. Chequear los rangos de las variables

En una reunión con los directivos se ha verificado la descripción del conjunto de datos y se ha concluido que las variables incluyen todos los casos posibles.

Tarea 7. Verificar otras representaciones de las variables.

Luego de una reunión con los usuarios se ha tomado la decisión de categorizar la variable "horario" en los valores "mañana" cuando la hora del turno es menor a las 13hs y "tarde" cuando es mayor.

4.4. Modelado

4.4.1. Modelado con CRISP-DM

CRISP-DM propone cuatro tareas para la fase de modelado: seleccionar la técnica de modelado, diseñar las pruebas del modelo, construir el modelo y evaluar el modelo.

Tarea 1 Seleccionar la técnica de modelado

Técnica de modelado

Debido a la naturaleza del problema y de las variables en estudio, se han seleccionado las siguientes técnicas de clasificación (la descripción de cada una se ubica en el anexo de este trabajo):

- Árboles de Decisión
- Regresión Logística
- Vecino más próximo (Knn)
- Clasificador de Naive Bayes

Para la construcción del modelo predictivo se utilizarán las variables del conjunto de datos final (Tabla 15).

Supuestos del modelo

No son necesarios supuestos para los modelos con los que se trabajará. Para aquellos modelos de regresión logística con más de una variable regresora cuantitativa se verificará la ausencia de multicolinealidad.

Tarea 2. Diseño de las pruebas del modelo

Los datos de prueba del modelo serán construidos a partir de una muestra aleatoria formada por el 35% de las instancias (3083 instancias). El 65% restante se utilizará para entrenar el modelo.

Se evaluará la capacidad del modelo en función de su matriz de confusión y la matriz de costos representada en la Tabla 16.

	Clasificó ASISTIÓ	**Clasificó NO ASISTIÓ**
ASISTIÓ	0	1
NO ASISTIÓ	2	0

Tabla 16. Matriz de costos

Como se puede observar, el costo más alto existe cuando el paciente no asiste a la consulta, y el modelo predice que asistirá. El costo de que el paciente asista cuando el modelo predice que no asistirá es menor. En este caso el turno podrá ser aprovechado, aunque provocará una mayor cantidad de personas en la sala de espera.

Además del costo se evaluará la tasa de acierto, es decir, la proporción de instancias bien clasificadas por el modelo.

Tarea 3 Construir el modelo

Antes de comenzar con el modelado es importante tener en cuenta que en el problema actual, las clases no se encuentran balanceadas. Es decir, la cantidad de pacientes que no asisten a su consulta es mucho menor a la

cantidad de pacientes que si lo hacen (Fig.30). Este escenario es muy frecuente en problemas del mundo real, donde el objetivo del modelo predictivo es identificar principalmente a las instancias de la clase minoritaria [12,37].

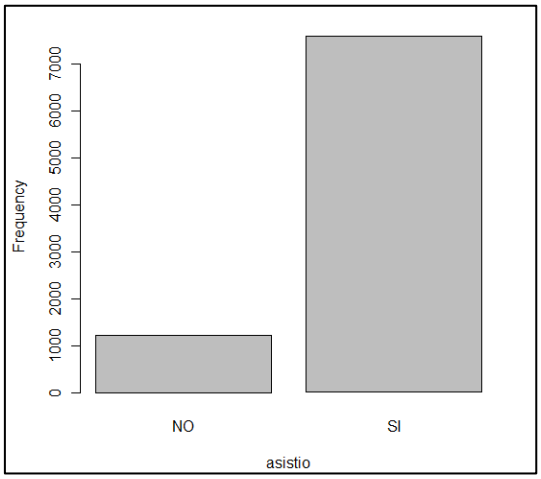

Fig. 30. Distribución de la variable explicada

Inicialmente, se comenzará trabajando con los datos originales. En caso de que los modelos no sean viables, se procederá a muestrear la clase mayoritaria.

Para cada técnica se realizarán distintas pruebas, seleccionando distintos subconjuntos de variables explicativas. Además, en caso de que el modelo lo permita, se evaluarán distintos valores de sus parámetros.

En este trabajo se muestran los mejores modelos obtenidos para cada técnica.

Arboles de decisión

Luego de utilizar el algoritmo J48 con el software WEKA, el árbol de menor costo es aquel que utiliza las variables "tiene HC", "es de la ciudad", "edad", "días antelación solicitud", "médico". El factor de confianza utilizado para la poda del árbol fue de 0,25 (valor por defecto). Mientras menor es este valor, se aumenta la poda de las ramas.

La Figura 31 muestra el árbol resultante.

Si bien la tasa de acierto es del 89% (Fig. 32), la matriz de confusión indica que de los 425 pacientes que no han asistido, 117 de ellos se han clasificado correctamente. Como es de esperarse, el modelo ha clasificado la mayor parte de las instancias dentro de la clase mayoritaria. El costo total del árbol es 308 x 2 + 30 x 1=646.

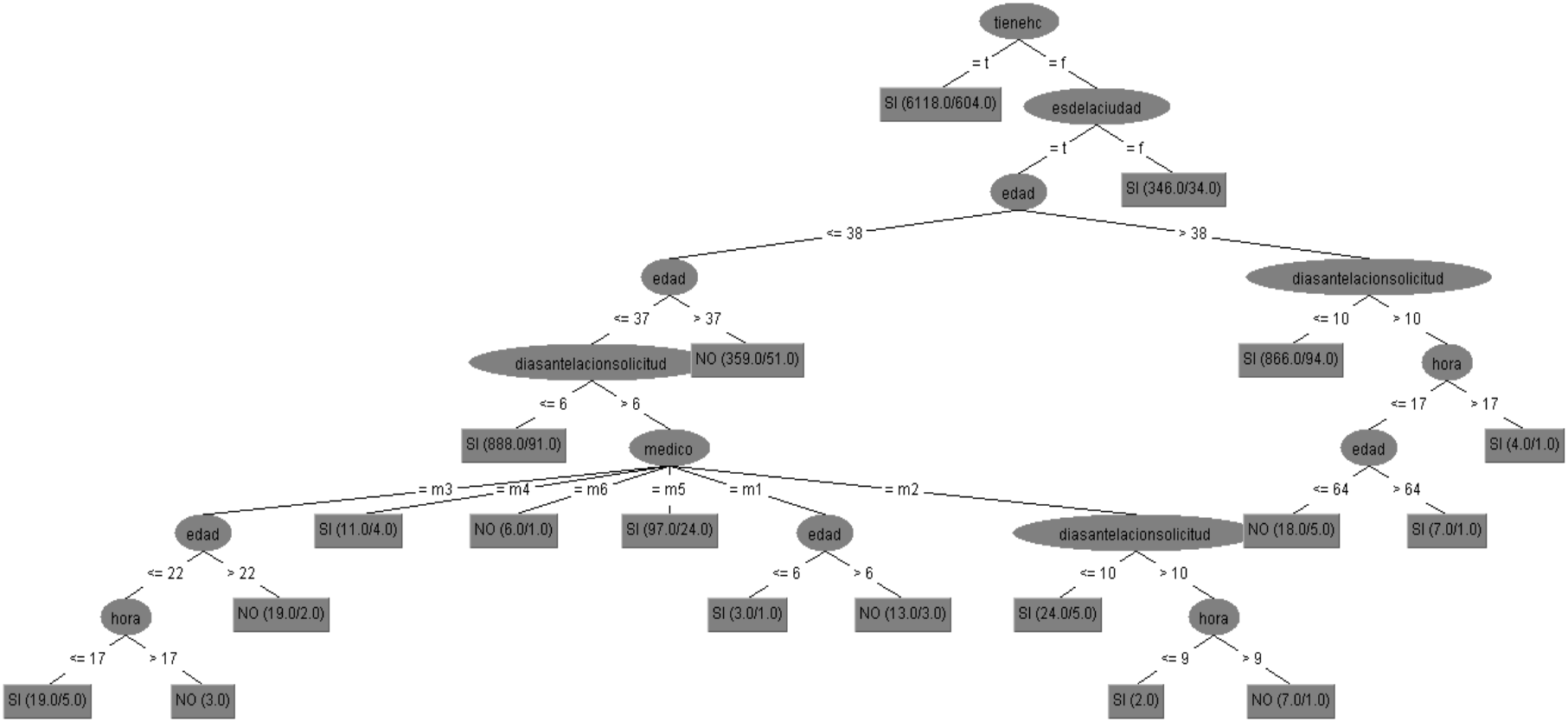

Fig. 31. Árbol generado por WEKA.

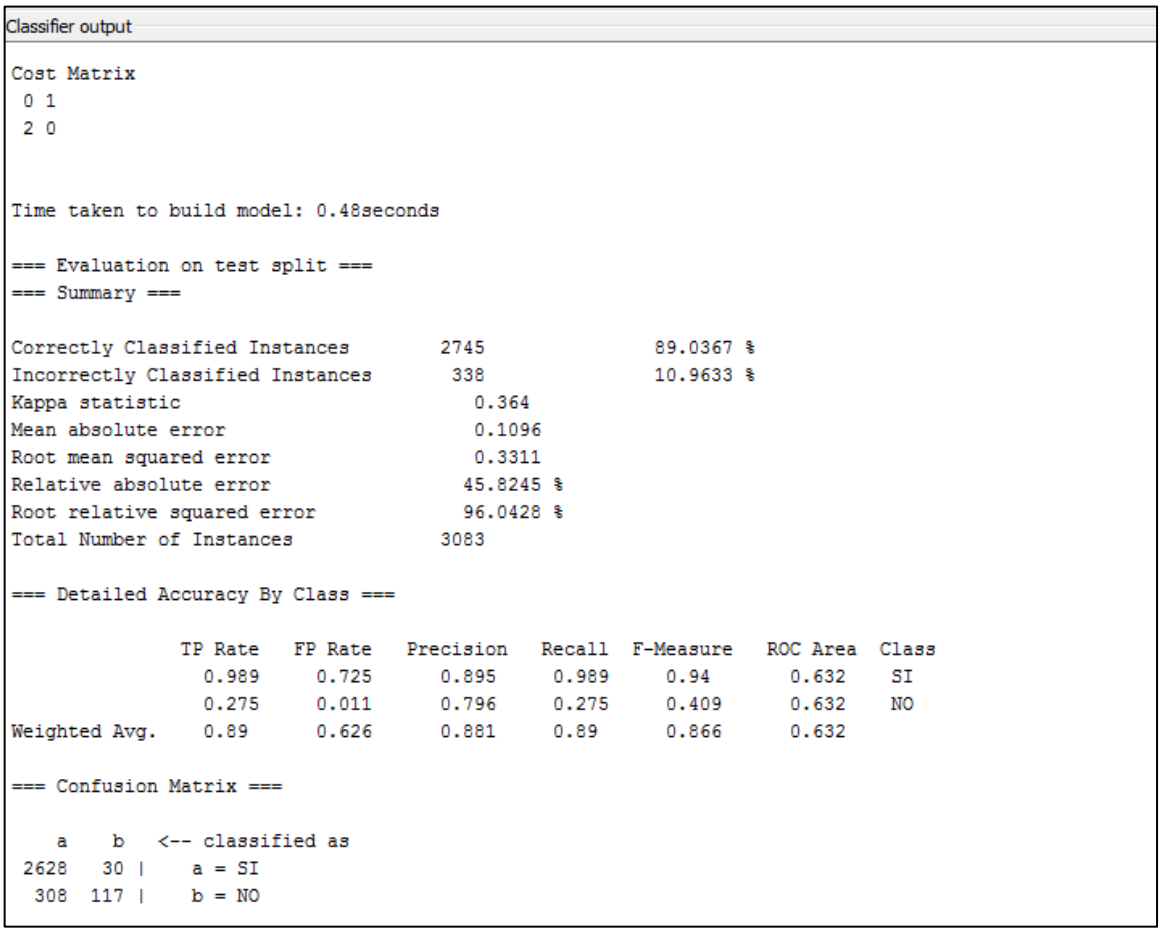

Fig. 32. Datos del árbol construido por WEKA.

Regresión logística

Para ajustar el modelo de regresión logística se ha recodificado la variable de respuesta.

La nueva variable de respuesta, llamada "ausente", recibe los valores 0 (cuando el paciente asistió) y 1 (cuando el paciente no asistió). Esta transformación se realiza porque en los modelos de regresión logística, es conveniente codificar la clase que interesa predecir con el valor 1.

El mejor modelo obtenido con el software R se representa en la Figura 33.

Las variables "días antelación solicitud", "es de la ciudad", "es sobreturno" y "tiene HC" resultan significantes para la regresión, con un p-valor menor a 0,05.

Como las predicciones para las instancias de prueba son una probabilidad (probabilidad de que el paciente se ausente en la consulta), se redondea:

- Si la probabilidad de que el paciente esté ausente es menor a 0,5: se toma 0 (asistirá).
- Si la probabilidad de que el paciente esté ausente es mayor a 0,5: se toma 1 (no asistirá).

```
> summary(modeloLogistico)

Call:
glm(formula = ausente ~ diasantelacionsolicitud + esdelaciudad +
    essobreturno + tienehc, family = binomial(logit), data = DatosEntrenamiento)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6723  -0.5651  -0.4481  -0.3954   2.7778

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.406274   0.193363 -12.444  < 2e-16 ***
diasantelacionsolicitud  0.087275   0.009156   9.532  < 2e-16 ***
esdelaciudad[T.t]      0.989804   0.191942   5.157 2.51e-07 ***
essobreturno[T.t]     -0.337339   0.162561  -2.075    0.038 *
tienehc[T.t]          -1.093233   0.081044 -13.489  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4591.8  on 5726  degrees of freedom
Residual deviance: 4325.0  on 5722  degrees of freedom
AIC: 4335

Number of Fisher Scoring iterations: 5
```

Fig. 33. Modelo de regresión logística obtenido con R.

La matriz de confusión (Fig.34) muestra que la mayor parte de las instancias fueron clasificadas con el valor 0 (el paciente asistió), no pudiendo predecir correctamente a los pacientes ausentes. El costo total del modelo es 435 x 2 + 4 x 1=847.

```
> table(DatosPrueba$ausente,DatosPrueba$predichosRedondeado)

      0    1
0 2642    4
1  435    2
```

Fig. 34. Matriz de confusión del modelo de regresión

Vecinos más próximos (KNN)

Se realizaron modelos con valores de k entre 1 y 10. El mejor modelo obtenido fue para un valor de k=2, con una tasa de acierto del 73%.

```
=== Confusion Matrix ===

   a    b   <-- classified as
2099  559 |    a = SI
 248  177 |    b = NO
```

Fig. 35. Matriz de confusión con el método KNN.

La matriz de confusión (Fig. 35) indica que de los pacientes que no han asistido (425), 177 de ellos se han clasificado correctamente (41%). El costo total del modelo es 248 x 2 + 559 x 1=1055.

Naive Bayes

En el mejor modelo obtenido con esta técnica se han utilizado todas las variables predictoras. Para aquellas que son numéricas, se utilizó una función del software WEKA que permite estimar su función de densidad.

La Figura 36 muestra la matriz de confusión luego de aplicar el método de Naive Bayes con WEKA. Aunque la tasa de acierto es buena (86%), se puede observar que los pacientes que realmente no asistieron no se han clasificado correctamente. El costo total del modelo es 304 x 2 + 111 x 1=719.

```
=== Confusion Matrix ===

   a    b   <-- classified as
2547  111 |    a = SI
 304  121 |    b = NO
```

Fig. 36. Matriz de confusión con el método Naive Bayes.

Tarea 4. Evaluar el modelo

Evaluación de los modelos

Los modelos obtenidos han demostrado una baja capacidad predictiva para pacientes que no asisten a su consulta. Como se puede ver en las matrices de confusión, a pesar de utilizar un análisis sensitivo al costo, la mayor parte de las instancias de prueba se clasifican en la clase mayoritaria ("asistió=si").

Es decir, que a pesar de que las tasas de acierto sean altas, las matrices de confusión indican que los modelos no son adecuados para identificar los casos que son de interés.

Dada esta situación, se procede a construir nuevamente los modelos, utilizando un muestreo de la clase mayoritaria.

Tarea 3. Construir el modelo (con muestreo de clase mayoritaria)

Un muestreo de la clase mayoritaria puede ayudar a reducir el desbalance existente en la variable de respuesta [12]. En el conjunto de datos original, la proporción de instancias es casi 6 a 1 (la clase mayoritaria supera en más de un 600% a la minoritaria).

Se procede entonces a realizar un muestreo de la clase mayoritaria (pacientes que asistieron), para reducir la proporción hasta 2 a 1.

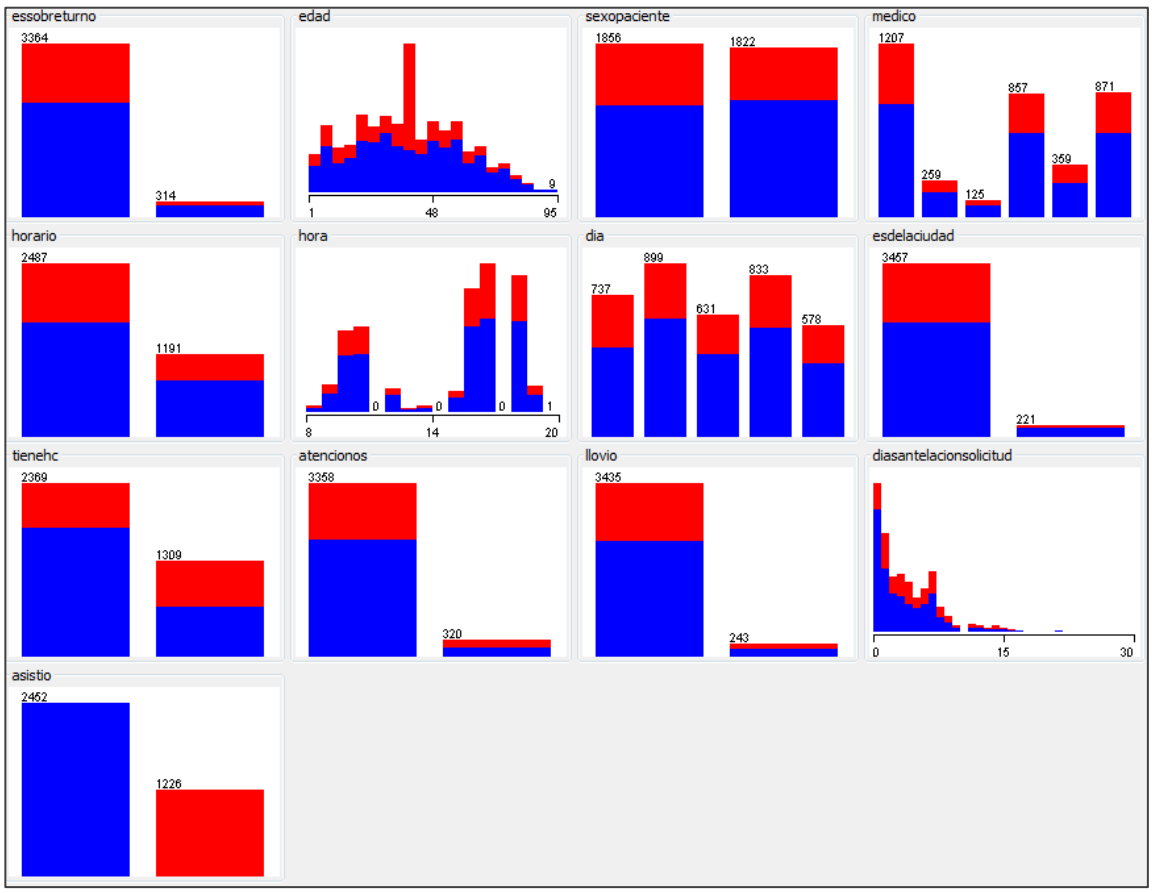

Fig. 37. Distribución del conjunto de datos muestreado.

Luego del muestreo la cantidad de casos se ha reducido a 3678 y la clase mayoritaria ahora tiene 2452 instancias (Fig.37).

También se ha modificado la matriz de costos, donde se ha aumentado en 0.5 el error al clasificar incorrectamente a un paciente que no ha asistido.

	Clasificó ASISTIÓ	**Clasificó NO ASISTIÓ**
ASISTIÓ	0	1
NO ASISTIÓ	2.5	0

Tabla 17. Matriz de costos modificada.

Árboles de decisión

El mejor modelo obtenido fue aquel formado por las variables "médico", "tiene HC", "es sobreturno", "días antelación solicitud", "edad", "sexo", "horario" y "atención OS". Con el objetivo de simplificar la estructura del árbol, se redujo el factor de confianza para la poda a 0,02. Esto genera que el algoritmo aumente la acción de poda, sin sacrificar en este caso la tasa de error.

El modelo obtenido logró clasificar mejor a los pacientes ausentes que en la etapa anterior (300/464, es decir, 65% bien clasificados), con una tasa de acierto general del 68% (Fig.38). El costo total del modelo es 164 x 2.5 + 247 x 1=657. La Figura 39 representa la estructura del árbol.

```
Cost Matrix
 0    1
 2.5  0


Time taken to build model: 0.13seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         876               68.0653 %
Incorrectly Classified Instances       411               31.9347 %
Kappa statistic                          0.3334
Mean absolute error                      0.4151
Root mean squared error                  0.4686
Relative absolute error                 92.3553 %
Root relative squared error             97.2317 %
Total Number of Instances             1287

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.7       0.353      0.778     0.7       0.737      0.727    SI
                 0.647     0.3        0.548     0.647     0.593      0.727    NO
Weighted Avg.    0.681     0.334      0.695     0.681     0.685      0.727

=== Confusion Matrix ===

   a   b   <-- classified as
 576 247 |   a = SI
 164 300 |   b = NO
```

Fig. 38. Salida del software WEKA para el modelo de árbol sobre los datos muestreados.

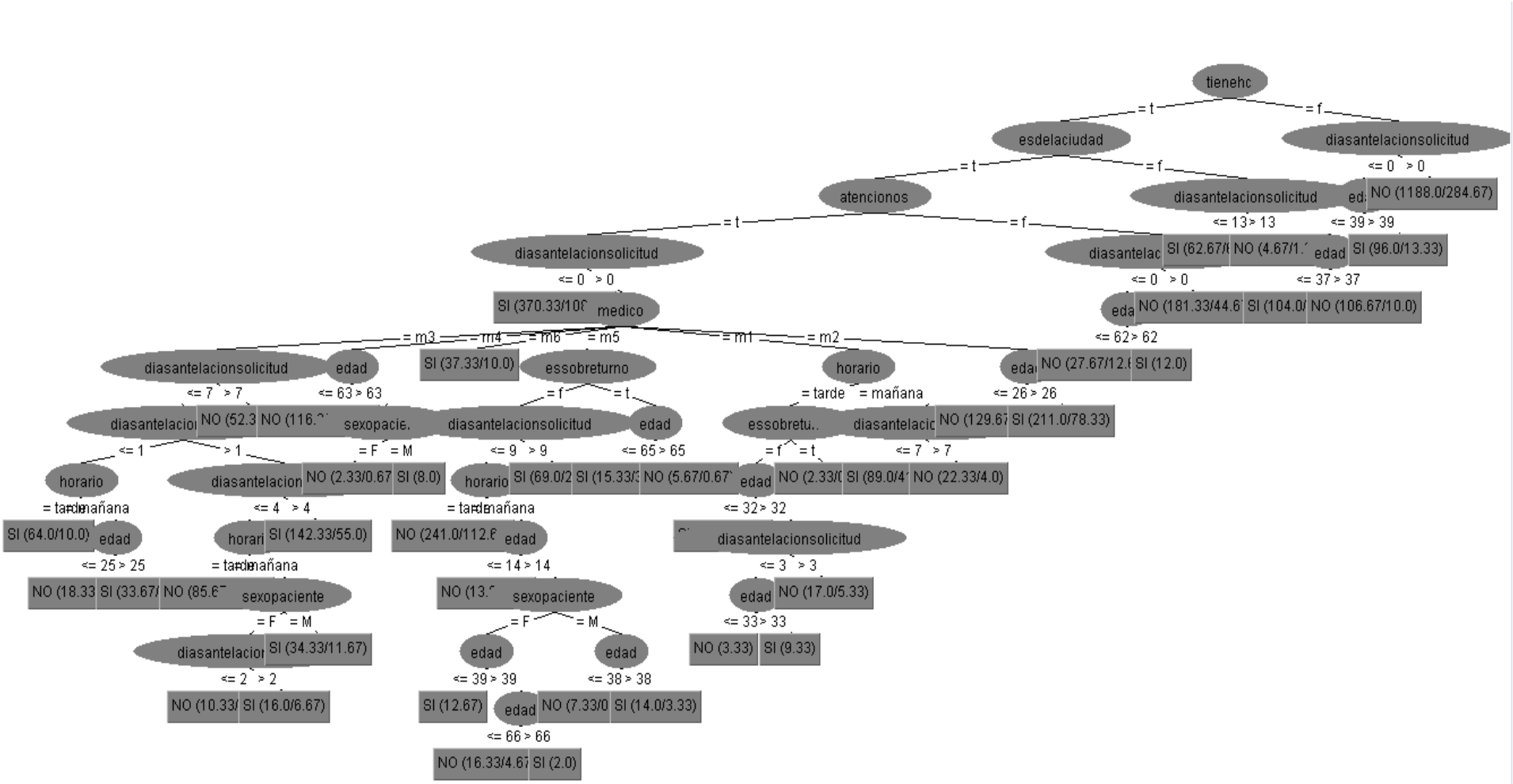

Fig. 39. Árbol obtenido con el software WEKA sobre los datos muestreados.

Regresión logística

El mejor modelo obtenido fue aquel formado por las variables explicativas "atención OS", "días antelación solicitud", "edad", "es de la ciudad", "es sobreturno" y "tiene HC" (Fig.40). El coeficiente de correlación entre los regresores "edad" y "días antelación solicitud" es de 0.07, por lo que se descarta multicolinealidad.

```
> summary(GLM.16)

Call:
glm(formula = ausente ~ atencionos + diasantelacionsolici + edad +
    esdelaciudad + essobreturno + tienehc, family = binomial(logit),
    data = DatosModelo)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8574  -0.8546  -0.6712   1.1860   2.1680

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.474384   0.285964  -1.659   0.0971 .
atencionos[T.t]      -0.734948   0.157378  -4.670 3.01e-06 ***
diasantelacionsolici  0.079179   0.011810   6.705 2.02e-11 ***
edad                 -0.005511   0.002327  -2.368   0.0179 *
esdelaciudad[T.t]     1.002710   0.224795   4.461 8.17e-06 ***
essobreturno[T.t]    -0.431133   0.191331  -2.253   0.0242 *
tienehc[T.t]         -1.054125   0.094994 -11.097  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2993.0  on 2390  degrees of freedom
Residual deviance: 2783.9  on 2384  degrees of freedom
AIC: 2797.9

Number of Fisher Scoring iterations: 4
```

Fig. 40. Modelo de regresión logística sobre los datos muestreados.

El modelo tiene una tasa de acierto general del 67%, y una tasa de acierto para pacientes ausentes del 22% (Fig.41). El costo total del modelo es 361 x 2.5 + 59 x 1=961.5.

```
> table(DatosPrueba$ausente,DatosPrueba$predichos)

      0   1
  0 764  59
  1 361 103
```

Fig. 41. Matriz de confusión para el modelo de regresión logística sobre datos muestreados

Vecinos más próximos (KNN)

Mediante esta técnica, el mejor valor obtenido fue con un k=5, con una tasa de acierto general del 59% y una tasa de acierto para pacientes ausentes de 62% (Fig.42).

El costo total del modelo es 163 x 2.5 + 352 x 1=784.5.

```
Classifier Model
IB1 instance-based classifier
using 5 nearest neighbour(s) for classification


Cost Matrix
 0    1
 2.5  0

Time taken to build model: 0seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         762               59.2075 %
Incorrectly Classified Instances       525               40.7925 %
Kappa statistic                          0.184
Mean absolute error                      0.4376
Root mean squared error                  0.5206
Relative absolute error                 97.3691 %
Root relative squared error            108.0228 %
Total Number of Instances             1287

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.572     0.373      0.731     0.572      0.642      0.637    SI
                 0.627     0.428      0.453     0.627      0.526      0.636    NO
Weighted Avg.    0.592     0.393      0.631     0.592      0.6        0.637

=== Confusion Matrix ===

   a   b   <-- classified as
 471 352 |   a = SI
 173 291 |   b = NO
```

Fig. 42. Resultados del método KNN sobre el conjunto de datos muestreado.

Naive Bayes

Con el clasificador de Naive Bayes, también se han obtenido mejores resultados que en el conjunto de datos sin muestrear (Fig. 43).

La tasa de acierto general del modelo es de 66%. La tasa de acierto para los pacientes ausentes es del 58%.

El costo total del modelo es 192 x 2.5 + 247 x 1=727.

```
Cost Matrix
 0    1
 2.5  0


Time taken to build model: 0.02seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         848               65.8897 %
Incorrectly Classified Instances       439               34.1103 %
Kappa statistic                          0.2789
Mean absolute error                      0.4229
Root mean squared error                  0.4681
Relative absolute error                 94.0865 %
Root relative squared error             97.1287 %
Total Number of Instances             1287

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.7       0.414      0.75      0.7       0.724      0.704    SI
                 0.586     0.3        0.524     0.586     0.553      0.704    NO
Weighted Avg.    0.659     0.373      0.669     0.659     0.663      0.704

=== Confusion Matrix ===

   a   b   <-- classified as
 576 247 |   a = SI
 192 272 |   b = NO
```

Fig. 43. Método de Naive Bayes sobre los datos muestreados.

Tarea 4. Evaluar el modelo (con muestreo de la clase mayoritaria)

Evaluación de los modelos

Los modelos obtenidos a partir del conjunto de datos muestreado resultaron mucho mejores que aquellos construidos sobre los datos originales.

La tasa de acierto para los pacientes ausentes ha aumentado considerablemente, ya que la proporción entre la clase mayoritaria y la minoritaria se ha reducido.

El mejor modelo que se obtuvo sobre el conjunto de datos muestreado fue utilizando árboles de decisión, con una tasa de acierto general del 68%, y una del 65% para pacientes ausentes. El modelo resultó tener el menor costo, con un valor de 657. Los modelos de KNN y Naive Bayes también han demostrado ser buenos, aunque su costo resultó mayor y la tasa de acierto para pacientes ausentes levemente menor.

4.4.2. Modelado con Catalyst

La metodología Catalyst propone las actividades de modelado en la fase "Selección de herramientas y modelado inicial".

Tarea 1. Estructurar los datos para el proceso

Catalyst propone dividir el conjunto de datos en tres partes: entrenamiento, prueba y evaluación. Los dos primeros se usan para construir el modelo, mientras que el último proporciona una confirmación de la capacidad predictiva del mismo.

En este caso, se ha tomado una nueva muestra de 2570 turnos correspondientes al periodo 01/08/2011 al 30/09/2011 para armar un conjunto de datos de evaluación que se utilizará en la siguiente fase (evaluación).

Tarea 2. Caracterizar las variables de entrada y salida

En este caso, las 10 variables de entrada son de distinto tipo (categóricas y numéricas). La variable de respuesta (output) es de tipo booleana (si/no). Las variables de entrada se agrupan en distintas combinaciones que se proporcionarán como entradas a los algoritmos.

Las variables fueron descriptas en la Tabla 15.

Tarea 3. Seleccionar una técnica de modelado supervisado

En este caso, como se mencionó en CRISP-DM, se trabajará con distintos algoritmos de clasificación: árboles de decisión, regresión logística, Naive Bayes y Vecinos más próximos (KNN).

Tarea 4. Crear el modelo de clasificación

En esta fase se lleva a cabo la aplicación de los algoritmos de minería. Esta actividad fue realizada en la tarea 4 dentro de la fase de Modelado en CRISP-DM.

Como las clases del conjunto de datos se encuentran desbalanceadas, la metodología sugiere la utilización de técnicas de sobre muestreo de la clase minoritaria. Esta es una alternativa al muestreo de la clase mayoritaria, decisión que se adoptó con CRISP-DM. Más allá de la técnica de balanceo que se utilice, el autor (Dorian Pyle) señala la importancia de tener equilibradas la cantidad de instancias entre ambas clases, para que el modelo de clasificación funcione adecuadamente.

Por cuestiones de acotación del estudio, se mantendrá la decisión adoptada en CRISP-DM, balanceando el conjunto de datos con técnicas de muestreo de la clase mayoritaria.

4.5. Evaluación

4.5.1 Evaluación con CRISP–DM

CRISP-DM lleva a cabo una primera evaluación técnica del modelo en la

fase anterior (Modelado). En esta etapa, se realiza una evaluación técnica final, una evaluación en función de los objetivos del negocio, y una revisión del proceso.

Tarea 1. Evaluar los resultados

Evaluación de los resultados obtenidos

Al aplicar las diferentes técnicas de clasificación sobre el conjunto de datos con todas las observaciones no se ha encontrado un modelo con una buena capacidad predictiva, debido al gran desbalance de las clases en la variable de respuesta ("asistió"). Este problema fue abordado con técnicas de muestreo de la clase mayoritaria, donde se redujo en gran medida la desproporción existente.

El conjunto de datos muestreado permitió a los algoritmos de clasificación mejorar la tasa de acierto para pacientes que no asistieron a su consulta (clase más importante para el problema en estudio).

El mejor modelo obtenido fue con árboles de decisión, el cual resultó con una tasa de acierto del 68%. Este valor cumple con el criterio de éxito establecido para el proyecto.

Sin embargo, es importante destacar que este modelo podría ser notablemente mejorado si se dispusiera de más variables. Información como el estado civil del paciente, cantidad de hijos, ocupación y otras cualidades referentes al paciente y al turno podrían reducir la tasa de error del modelo.

Modelos evaluados y aprobados

El modelo aprobado resultante es un árbol de decisión, construido a partir de las variables "médico", "tiene hc", "es sobreturno", "días antelación solicitud", "edad", "sexo", "horario" y "atención por obra social". El factor de confianza para la poda es de 0,02. La matriz de costos está representada en la Tabla 17.

El modelo tiene una tasa de acierto general del 68%.

Tarea 2. Revisión del proceso

El proyecto de minería de datos no ha presentado mayores inconvenientes en las etapas de análisis del problema y preparación de los datos. Sin embargo, la fase de modelado resultó dificultosa por el desbalanceo en las clases de la variable de respuesta. Se han producido una gran cantidad de modelos, donde se priorizó la capacidad de cada uno para predecir a los pacientes ausentes.

Finalmente, se detectó que con un muestreo de la clase mayoritaria, una técnica específica (árboles de decisión) y ciertos valores en la matriz de costos, se pudo obtener un modelo aceptable para el problema en estudio.

No hay actividades importantes que hayan sido omitidas. Sería importante tener en cuenta la experiencia obtenida con este conjunto de datos desbalanceado para futuros proyectos.

Tarea 3. Determinar las próximas etapas

Se ha tomado la decisión de implementar el modelo obtenido en el sistema operacional de turnos. Con esta información el recepcionista podrá confirmar telefónicamente la asistencia de aquellos pacientes con alta probabilidad de ausentarse, además de otorgar sobreturnos en función de las predicciones del modelo.

4.5.2. Evaluación con Catalyst

Las actividades de evaluación se llevan a cabo en la fase "Refinar el modelo" de Catalyst. En ella se propone hacer una evaluación técnica, chequeando la matriz de confusión, y una del negocio, revisando que los objetivos se hayan cumplido.

Tarea 1. Verificar la matriz de confusión

La matriz de confusión para los datos de prueba revela que el árbol ha clasificado erróneamente un total de 411 turnos, de 1287 (Fig.38). El modelo logró clasificar correctamente al 65% de los pacientes ausentes, con una tasa de acierto general del 68%.

Catalyst propone también chequear el modelo final con un conjunto de datos llamado "evaluación". Este conjunto de datos se ha creado en la fase anterior, y corresponde a un total de 2570 turnos, correspondientes al periodo 01/08/2011 al 30/09/2011.

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     datosEvaluacion-weka.filters.unsupervised.attribute.Remove-R6-7
Instances:     unknown (yet). Reading incrementally
Attributes:   10

=== Summary ===

Correctly Classified Instances        1568               61.0117 %
Incorrectly Classified Instances      1002               38.9883 %
Kappa statistic                          0.1572
Mean absolute error                      0.4367
Root mean squared error                  0.4941
Total Number of Instances             2570

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.592     0.265      0.938     0.592      0.726      0.724    SI
                 0.735     0.408      0.208     0.735      0.325      0.724    NO
Weighted Avg.    0.61      0.283      0.845     0.61       0.675      0.724

=== Confusion Matrix ===

    a    b   <-- classified as
 1327  915 |    a = SI
   87  241 |    b = NO
```

Fig. 44. Desempeño del modelo con los datos de evaluación.

Los resultados obtenidos (Fig. 44) indican que el modelo ha podido identificar a 241 de 328 pacientes ausentes (73%), con una capacidad predictiva general de 61%. Se puede confirmar entonces que el modelo sigue siendo válido en instancias diferentes a las que se utilizaron para construirlo.

Tarea 2. Comprobación de la validez antes de la implementación

En esta tarea se revisan los requerimientos de la implementación y los requerimientos del negocio, ambos planteados en la fase de Modelado del Negocio.

En el caso de estudio, el árbol de decisión logra predecir con un 68% la asistencia del paciente a su consulta. Con esta información la recepcionista podrá regular la cantidad de sobreturnos y corroborar con los pacientes potencialmente ausentes su asistencia a la cita. El modelo obtenido cumple con las expectativas manifestadas por el personal interesado.

4.6. Implementación

4.6.1. Implementación con CRISP-DM

CRISP-DM lleva a cabo la fase de implementación mediante cuatro tareas: planificación de la implementación, planificación del monitoreo, reporte final y revisión del proyecto.

Tarea 1, Planificar la implementación

Plan de implementación

La implementación del modelo obtenido se realizará mediante las siguientes tareas:

1. Dar a conocer los resultados del proyecto a los directivos, personal de administración y recepcionistas. Para ello se procederá a crear un informe final con un resumen del proyecto.
2. Discutir con los directivos los resultados. Tener en cuenta la utilidad del conocimiento adicional obtenido (además del modelo) y de nuevas hipótesis que puedan surgir para futuras investigaciones.
3. En base al modelo seleccionado, incorporar al sistema de informático de turnos funciones para predecir que pacientes probablemente no asistan a su consulta.
4. Capacitar a los usuarios finales del sistema. Aclarar que las predicciones están sujetas a una probabilidad.
5. Luego de un cierto periodo, verificar si la tasa de ausentismo ha disminuido (objetivo organizacional).

Tarea 2. Planificar el mantenimiento y monitoreo

Plan de mantenimiento y monitoreo

Periódicamente se realizarán pruebas sobre la tasa de acierto del modelo, para verificar que se encuentre dentro de los márgenes de tolerancia (aproximadamente el 70% de los pacientes ausentes debe ser detectado).

Como se enuncia en el punto 9 de la fase Modelado del Negocio en Catalyst, el modelo podría dejar de ser válido cuando:

- Se contraten más médicos.
- Factores externos a la organización provoquen un cambio en el hábito de asistencia de los pacientes a su turno.

En estos casos, se requerirá probar nuevamente el modelo bajo las nuevas condiciones para validar su vigencia.

Tarea 3. Crear un reporte final

Reporte final

El proyecto de minería de datos se ha desarrollado exitosamente, no sólo por los modelos descubiertos, sino por toda la información adyacente que se ha generado, que aporta material valioso a la organización.

El personal de la organización ha colaborado en todo momento con el proyecto, participando activamente del mismo.

Los datos no han presentado problemas en cuanto a su disponibilidad y veracidad. Salvo la variable "edad", todos los campos estaban completos y no presentaban datos ausentes.

El modelo que se ha seleccionado para la fase de implementación fue un árbol de decisión, construido a partir de las variables "médico", "tiene HC", "es sobreturno", "días antelación solicitud", "edad", "sexo", "horario" y "atención OS". La variable de respuesta, "asistió", constituye la salida del árbol.

Durante el modelado, se realizó especial énfasis en la tasa de acierto del modelo para los pacientes ausentes, acción que se ve reflejada en las penalizaciones de la matriz de costos.

Es importante destacar los inconvenientes que trajo a la fase de modelado el desbalanceo de las clases de la variable de respuesta. A pesar de haber obtenido un modelo aceptable tras muestrear la clase mayoritaria, el mismo podría mejorarse notablemente si se dispusiera de mayor cantidad de variables.

El modelo se implementará en el sistema de gestión de turnos de la clínica, y será monitoreado periódicamente para comprobar su vigencia.

Tarea 4. Revisión del proyecto

Revisión del proyecto

El proyecto se ha llevado a cabo en forma exitosa, en el tiempo programado con el cliente. Se ha logrado una alta participación de todo el personal involucrado en el proyecto.

Los datos que disponía el cliente pudieron ser explorados y explotados, obteniendo un buen modelo predictivo con la técnica de árboles de decisión.

Se han producido retrasos en la fase de modelado, ya que el equipo de trabajo tenía poca experiencia en materia de conjuntos de datos desbalanceados. Esta situación sirvió como aprendizaje para casos futuros.

4.6.2. Implementación con Catalyst

La fase de implementación en Catalyst consiste básicamente en revisar y corregir, en caso de ser necesario, el plan de implementación creado en la fase de Modelado del Negocio.

Tarea 1. Revisar los requerimientos de la implementación

Los requerimientos de la implementación ya fueron definidos inicialmente en la tarea 9 de la fase "Modelado del negocio".

Se han revisado los mismos con el personal interesado y no han surgido modificaciones.

Tarea 2. Preparar la explicación del modelo

En el caso de estudio se realizará una reunión con los directivos y personal involucrado en el proyecto, para revisar:

- Problemas que dieron origen al proyecto.
- Datos que se utilizaron.
- Modelo obtenido.
- Forma en la que será implementado (Plan de implementación). Este punto coincide con la tarea 1 de la fase Implementación en CRISP-DM.
- Discusión sobre potenciales mejoras, como la recolección de más datos del paciente para futuros modelos.

5. Comparación de las metodologías CRISP-DM y Catalyst

Ha llegado el momento de confrontar las metodologías CRISP-DM y Catalyst. El marco desarrollado en el Capítulo 3 será la herramienta que explicitará los puntos a comparar, los cuales serán evaluados en forma positiva o negativa.

Se recuerda la posibilidad de crear una ponderación para las características que integran el marco comparativo, quedando la valoración de las mismas al criterio del usuario de esta herramienta.

Durante la evaluación se contabilizará para cada aspecto las valoraciones positivas sobre el total de características evaluadas. Es importante tener en cuenta que los resultados podrían variar sensiblemente si se trabajara con un sistema de puntajes.

5.1. Evaluación del nivel de detalle en las actividades de cada fase

Característica	CRISP-DM	Catalyst	Comentario
1.1. ¿Se definen actividades específicas para cada fase del proceso?	SI	SI	
1.2. ¿Se explicitan los pasos a seguir para llevar a cabo cada actividad?	SI	SI	Catalyst explica cómo llevar a cabo cada actividad referenciando al contenido del libro donde la misma fue publicada [26]. CRISP-DM indica breves instrucciones en la sección "Guía del usuario".
1.3. ¿Se definen las entradas de cada actividad?	NO	NO	En ninguna de las dos metodologías se explicitan las entradas (como por ejemplo los entregables que sirven de fuente) para cada actividad.
1.4. ¿Se definen las salidas de cada actividad?	SI	SI	
1.5. ¿Se provee una guía de buenas prácticas para cada una de las actividades específicas?	SI	SI	En ambas metodologías se proponen consejos sobre la ejecución de cada actividad.
Valoraciones positivas **CRISP-DM: 4/5 = 80%** **CATALYST: 4/5 = 80%**			

Tabla 18. Evaluación del nivel de detalle en las actividades que componen cada fase

En este aspecto del marco comparativo ambas metodologías han obtenido el mismo porcentaje de valoraciones positivas (80%). La única característica que ninguno de los dos enfoques ha cumplido es la definición de las entradas para cada actividad.

5.2. Evaluación de los escenarios de aplicación

Característica	CRISP-DM	Catalyst	Comentario
2.1. ¿Se especifican actividades para la definición y el análisis del problema u oportunidad con el cual colaborará la minería de datos?	SI	SI	
2.2. ¿Se consideran puntos de partida alternativos donde el usuario no refiere un problema sino que sólo desea explorar sus datos?	NO	SI	En Catalyst se consideran cinco escenarios posibles de partida para el proyecto, incluyendo aquel donde el usuario manifiesta que sólo desea explorar sus datos.
2.3. ¿La metodología es independiente del dominio de aplicación?	SI	SI	
2.4. ¿La metodología es aplicable a proyectos de diferente tamaño?	SI	SI	
Valoraciones positivas **CRISP-DM: 3/4 = 75%** **CATALYST: 4/4 = 100%**			

Tabla 19. Evaluación de los escenarios de aplicación.

En cuanto a los escenarios de aplicación, Catalyst ha demostrado ser la metodología más completa ya que cumple con el 100% de las características en este aspecto.

CRISP-DM ha logrado cumplir tres de cuatro puntos de evaluación, obteniendo un puntaje negativo en la segunda característica por no considerar escenarios de partida alternativos en el proyecto.

5.3. Evaluación de las actividades específicas en cada fase

Análisis del problema

Característica	CRISP-DM	Catalyst	Comentario
3.1. ¿Se propone una evaluación general de la organización?	SI	SI	
3.2. ¿Se identifica al personal involucrado en el proyecto (stackeholders)?	SI	SI	
3.3. ¿Se define el problema u oportunidad de negocio?	SI	SI	

Tabla 20.a. Evaluación de las actividades en la fase de análisis del problema.

3.4. ¿Se propone una evaluación de las fuentes de datos?	NO	SI	CRISP-DM propone la evaluación de las fuentes de datos en la próxima fase, luego de realizar el plan del proyecto.
3.5. ¿Se analizan todas las soluciones posibles al problema?	NO	SI	Catalyst propone efectuar un análisis de todas las posibles soluciones al problema.
3.6. ¿Se especifican los objetivos del proyecto?	SI	SI	
3.7. ¿Se define un criterio de éxito para el proyecto?	SI	NO	En Catalyst no se propone la definición del criterio de éxito para el proyecto (tanto técnico como organizacional).
3.8. ¿Se realiza una evaluación general de las técnicas de minería que podrían utilizarse?	SI	NO	Sólo CRISP-DM propone la evaluación inicial de las técnicas de minería que podrían utilizarse en el proyecto.
3.9. ¿Se especifica de qué forma el usuario utilizará el nuevo conocimiento?	NO	SI	Catalyst propone realizar en etapas tempranas una planificación de la implementación.
Valoraciones positivas **CRISP-DM: 6/9 = 66%** **CATALYST: 7/9 = 77%**			

Tabla 20.b. Evaluación de las actividades en la fase de análisis del problema.

Como se puede observar en la tabla 20, la metodología Catalyst ha obtenido el mejor resultado, cumpliendo el 77% de las características evaluadas. Las únicas dos características que fueron negativas en Catalyst están presentes en CRISP-DM (definición de un criterio de éxito y evaluación de las técnicas de minería).

En esta fase de análisis del problema, Catalyst propone la obtención y el estudio de las fuentes de datos con las que se trabajará, acción que resulta de gran importancia para la planificación del proyecto. CRISP-DM propone armar un plan de proyecto solamente identificando las fuentes de datos, sin estudiar el esfuerzo de integración y el formato actual de las mismas.

Otra diferencia entre ambos enfoques surge cuando se realiza un análisis de soluciones al problema. Catalyst propone el estudio de todas las alternativas de solución, incluyendo la posibilidad de "no hacer nada", mientras que CRISP-DM asume que la minería de datos es la solución al problema.

Finalmente, Catalyst propone en esta etapa documentar la forma en la que el "nuevo conocimiento" se entregará y se difundirá. CRISP-DM pospone esta actividad para el final del proyecto.

Selección y preparación de los datos

Característica	CRISP-DM	Catalyst	Comentario
3.10. ¿Se propone un análisis exploratorio inicial de los datos?	SI	SI	
3.11. ¿Se sugieren actividades para la limpieza de los datos?	SI	SI	
3.12. ¿Se contemplan actividades para la transformación de variables y la creación de atributos derivados?	SI	SI	
3.13. ¿Se realiza un análisis descriptivo final sobre los datos depurados?	NO	NO	En ninguna de las dos metodologías se proponen actividades para el análisis descriptivo de las variables transformadas.
3.14. ¿Se verifica con el usuario la completitud del conjunto de datos final?	NO	SI	CRISP-DM no propone actividades para la revisión del conjunto de datos final con el usuario.
Valoraciones positivas **CRISP-DM: 3/5 = 60%** **CATALYST: 4/5 = 80%**			

Tabla 21. Evaluación de las actividades en la fase de selección y preparación de los datos.

En la fase de selección y preparación de los datos, Catalyst ha cumplido con el 80% de las características mientras que CRISP-DM lo ha hecho en un 60%. Ninguna de las dos metodologías propone explícitamente la realización de un análisis descriptivo sobre el conjunto de datos final o vista minable.

CRISP-DM no propone actividades de revisión con el usuario del conjunto de datos final, lo cual permitiría una validación general de la completitud, formato e interpretación de los datos.

Modelado

Característica	CRISP-DM	Catalyst	Comentario
3.15. ¿Se efectúa una selección de las técnicas que se utilizarán?	SI	SI	
3.16. ¿Se planifica la forma en la que se evaluarán los resultados?	SI	NO	En Catalyst no se proponen actividades para especificar la forma en la que el equipo de trabajo evaluará los modelos obtenidos.
3.17. ¿Se efectúa una evaluación inicial de los modelos obtenidos?	SI	SI	

Tabla 22.a. Evaluación de las actividades en la fase de modelado.

3.18. ¿Se proveen directivas para el caso donde se dificulta el descubrimiento de patrones?	NO	SI	Catalyst propone los pasos a seguir ante diferentes dificultades que pueden presentarse durante la fase de modelado.
Valoraciones positivas **CRISP-DM: 3/4 = 75%** **CATALYST: 3/4 = 75%**			

Tabla 22.b. Evaluación de las actividades en la fase de modelado.

En la Tabla 22 se puede apreciar que ambas metodologías cumplieron la misma cantidad de puntos de evaluación (75%).

Catalyst no propone actividades para planificar la forma en la que se evaluarán los patrones obtenidos. Por otro lado, CRISP-DM no propone directivas para el caso donde no se encuentren patrones en el conjunto de datos.

Evaluación

Característica	CRISP-DM	Catalyst	Comentario
3.19. ¿Se interpretan los modelos en función de los objetivos organizacionales?	SI	SI	
3.20. ¿Se comparan y ponderan los modelos obtenidos?	SI	SI	
3.21. ¿Se propone una revisión general del proceso?	SI	SI	
3.22. ¿Se proveen directivas para el caso donde ninguno de los modelos obtenidos resulta viable?	SI	SI	
Valoraciones positivas **CRISP-DM: 4/4 = 100%** **CATALYST: 4/4 =100%**			

Tabla 23. Confrontación de las actividades en la fase de evaluación.

En la fase de evaluación, ambas metodologías han logrado cumplir el 100% de las características.

Implementación

Característica	CRISP-DM	Catalyst	Comentario
3.23. ¿Se planifica la implementación del nuevo conocimiento?	SI	SI	
3.24. ¿Se propone la creación de un programa de mantenimiento?	SI	SI	

Tabla 24.a. Evaluación de las actividades para la fase de implementación.

3.25. ¿Se entrega al usuario un resumen del proyecto?	SI	SI	
3.26. ¿Se documenta la experiencia adquirida por el equipo de trabajo?	SI	NO	CRISP-DM propone la revisión del proyecto para documentar la experiencia adquirida en el transcurso del mismo.
Valoraciones positivas **CRISP-DM: 4/4 =100%** **CATALYST: 3/4 =75%**			

Tabla 24.b. Evaluación de las actividades para la fase de implementación.

En la fase de implementación, sólo CRISP-DM ha logrado cumplir todas las características de evaluación. Catalyst no propone actividades para la documentación de la experiencia adquirida por el equipo de trabajo a lo largo del proyecto.

Evaluación general de las actividades específicas que componen cada fase.

En la Tabla 25 se sintetizan los resultados de la evaluación de las actividades específicas que componen cada fase del proceso. Se ha calculado un total general para este aspecto, obteniendo ambas metodologías un resultado similar (CRISP-DM cumple el 77% de las características mientras que Catalyst el 80%).

Fase	CRISP-DM	Catalyst
Análisis del problema	6/9 (66%)	7/9 (77%)
Selección y preparación de los datos	3/5 (60%)	4/5 (80%)
Modelado	3/4 (75%)	3/4 (75%)
Evaluación	4/4 (100%)	4/4 (100%)
Implementación	4/4 (100%)	3/4 (75%)
Total	**20/26 (77%)**	**21/26 (80%)**

Tabla 25. Evaluación general de las actividades específicas.

5.4. Evaluación de las actividades para la dirección del proyecto

Gestión del alcance

Característica	CRISP-DM	Catalyst	Comentario
4.1. ¿Se propone la selección de los entregables que se generarán durante el proyecto?	SI	SI	
4.2. ¿Se especifican actividades de control del alcance?	NO	NO	
Valoraciones positivas **CRISP-DM: 1/2 = 50%** **CATALYST: 1/2 =50%**			

Tabla 26. Evaluación de la gestión del alcance.

Como se puede observar en la Tabla 26, ambas metodologías proponen actividades para la definición y planificación del alcance, pero no para el control del trabajo y entregables planificados.

Gestión del tiempo

Característica	CRISP-DM	Catalyst	Comentario
4.3. ¿Se realiza una definición y secuenciación de las actividades que se ejecutarán durante el proyecto?	SI	SI	
4.4. ¿Se realiza una estimación de la duración de cada actividad?	SI	SI	
4.5. ¿Se construye un cronograma para el proyecto?	SI	SI	
4.6. ¿Existen actividades de control del cronograma?	NO	NO	
Valoraciones positivas **CRISP-DM: 3/4 = 75%** **CATALYST: 3/4 =75%**			

Tabla 27. Evaluación de la gestión del tiempo.

En cuanto a la gestión del tiempo ambas metodologías cumplen el 75% de las características evaluadas, las cuales están relacionadas a actividades de planificación. El único punto ausente en ambos enfoques son las actividades de control del cronograma.

Gestión del costo

Característica	CRISP-DM	Catalyst	Comentario
4.7. ¿Se efectúa una estimación de los recursos afectados por cada actividad?	SI	SI	
4.8. ¿Se realiza una estimación de los costos del proyecto?	NO	NO	En ninguna de las dos metodologías se realiza una estimación de los costos del proyecto basada en el plan definido para el mismo.
4.9. ¿Se construye un presupuesto de costos?	NO	NO	
4.10. ¿Existen actividades de control del presupuesto a medida que avanza el proyecto?	NO	NO	
Valoraciones positivas **CRISP-DM: 1/4 = 25%** **CATALYST: 1/4 = 25%**			

Tabla 28. Evaluación de la gestión del costo.

La gestión del costo ha obtenido un porcentaje de valoraciones positivas muy bajo (25%) en ambas metodologías.

Tanto en CRISP-DM como en Catalyst, se propone un análisis costo–beneficio cuyo objetivo principal sería la justificación económica del proyecto. Sin embargo, no se propone la creación de un presupuesto formal basado en el cronograma, es decir, una estimación de los costos de los recursos necesarios para completar las actividades del mismo.

En ninguna de las dos metodologías se propone un adecuado plan de gestión del costo. Tampoco se proponen actividades de control del presupuesto, lo cual es lógico ya que este control debería realizarse sobre una planificación previa. Algunos investigadores están trabajando sobre este tema, como en [18] donde se propone un modelo de estimación del costo basado en el estándar COCOMO[7].

Gestión del equipo de trabajo

Característica	CRISP-DM	Catalyst	Comentario
4.11. ¿Se efectúa una planificación de los recursos humanos?	SI	SI	

Tabla 29.a. Evaluación de la gestión del equipo de trabajo.

[7] COCOMO (COnstructive COst MOdel) es un modelo matemático empírico utilizado para la estimación de costos en los proyectos de software.

4.12. ¿Se proponen actividades para motivar la interacción entre los miembros del equipo?	NO	NO	
4.13. ¿Se efectúa un seguimiento del rendimiento de los recursos humanos?	NO	NO	
Valoraciones positivas **CRISP-DM: 1/3 = 33%** **CATALYST: 1/3 = 33%**			

Tabla 29.b. Evaluación de la gestión del equipo de trabajo.

CRISP-DM y Catalyst han cumplido uno de tres puntos en la evaluación de la gestión del equipo de trabajo. Si bien en ambas metodologías se planifican los recursos humanos, no se proponen actividades que fomenten la integración grupal y no se efectúa un seguimiento del desempeño del personal.

Gestión del riesgo

Característica	CRISP-DM	Catalyst	Comentario
4.14. ¿Se efectúa una identificación de los riesgos del proyecto?	SI	SI	
4.15. ¿Se realiza una cuantificación de los riesgos?	NO	SI	Si bien CRISP-DM propone la identificación del riesgo, no hace referencia al cálculo de su probabilidad e impacto.
4.16. ¿Se planifican acciones de respuesta ante cada riesgo?	SI	NO	En Catalyst no se explicitan claramente actividades para la planificación de la respuesta al riesgo.
4.17. ¿Existen actividades de supervisión y control de los riesgos?	NO	NO	
Valoraciones positivas **CRISP-DM: 2/4 = 50%** **CATALYST: 2/4 = 50%**			

Tabla 30. Evaluación de la gestión del riesgo.

En la gestión del riesgo, CRISP-DM y Catalyst han cumplido el 50% de las características evaluadas.

CRISP-DM no propone claramente actividades para la cuantificación y priorización de los riesgos. Por su parte la metodología Catalyst no especifica la necesidad de elaborar planes de acción (proactivos y reactivos) para los riesgos identificados.

En ninguna de las dos metodologías se proponen actividades de control, como la reevaluación periódica de los riesgos ya identificados (para

determinar si han cambiado) y auditorías de control sobre las actividades proactivas y reactivas planificadas.

Evaluación general de las actividades para la dirección del proyecto.

La Tabla 31 muestra las valoraciones obtenidas por cada metodología en las diferentes áreas. Ambas han cumplido con el 47% de las características evaluadas, dejando en evidencia la falta de madurez de ambos enfoques en materia de dirección de proyectos. El área con menor madurez resultó ser la gestión del costo, con sólo el 25% de valoraciones positivas en ambas metodologías.

Área	CRISP-DM	Catalyst
Gestión del alcance	1/2 (50%)	1/2 (50%)
Gestión del tiempo	3/4 (75%)	3/4 (75%)
Gestión del costo	1/4 (25%)	1/4 (25%)
Gestión del equipo de trabajo	1/3 (33%)	1/3 (33%)
Gestión del riesgo	2/4 (50%)	2/4 (50%)
Total	**8/17 (47%)**	**8/17 (47%)**

Tabla 31. Evaluación general para las actividades de dirección del proyecto.

5.5. Evaluación final

En este capítulo se ha utilizado el marco comparativo propuesto en este trabajo de tesis para confrontar las metodologías CRISP-DM y Catalyst. La Tabla 32 resume los resultados obtenidos, donde para cada aspecto se han contabilizado las características positivas sobre el total de características evaluadas.

Aspecto	CRISP-DM	Catalyst
Nivel de detalle en la descripción de las actividades.	4/5 (80%)	4/5 (80%)
Escenarios de aplicación	3/4 (75%)	4/4 (100%)
Actividades específicas de cada fase	20/26 (77%)	21/26 (81%)
Actividades de dirección del proyecto	8/17 (47%)	8/17 (47%)
Total de características cumplidas	**35/52 (67%)**	**37/52 (71%)**

Tabla 32. Evaluación final de todos los aspectos del marco comparativo

La evaluación final demuestra que se ha obtenido un resultado muy parejo. La metodología Catalyst logró cumplir con 37 de las 52 características que componen el marco (es decir, el 71%). CRISP-DM ha obtenido un resultado levemente inferior cumpliendo el 67% de los puntos evaluados.

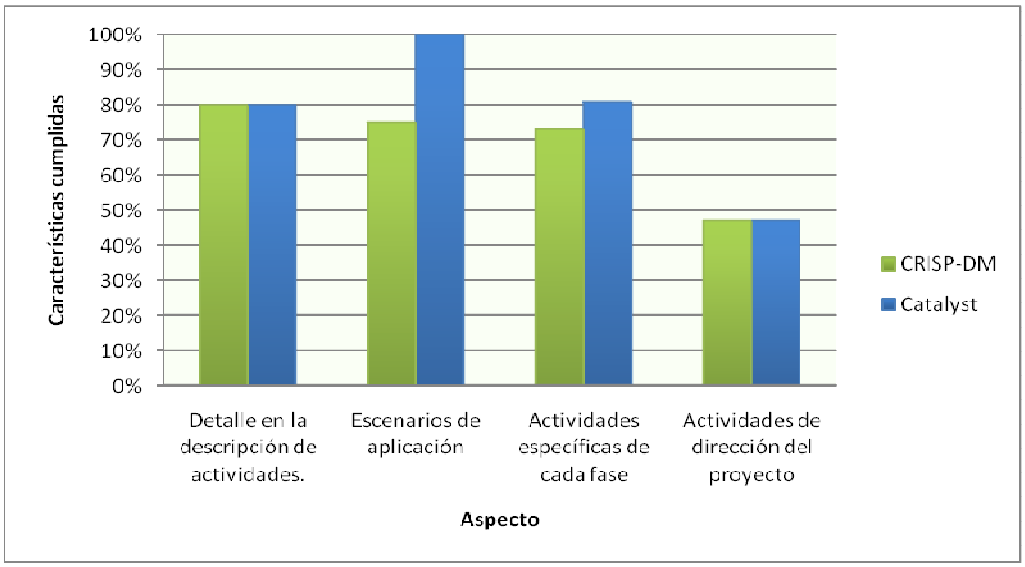

Fig. 45. Porcentaje de características presentes en cada metodología

Como se puede observar en la Figura 45, ambas metodologías han resultado con un puntaje similar en la mayoría de los aspectos, salvo en los escenarios de aplicación donde Catalyst ha logrado cumplir el 100% de las características evaluadas.

Se puede apreciar también que ninguna de las dos metodologías tuvo un buen desempeño en las actividades de dirección del proyecto, ya que ambas han obtenido un puntaje inferior al 50%, evidenciando la falta de madurez en este aspecto.

A partir de esta evaluación, se puede concluir que CRISP-DM y Catalyst reúnen un buen porcentaje de las características propuestas en este marco comparativo, aunque ambos enfoques deberían complementarse con actividades destinadas a la dirección del proyecto, especialmente en la gestión del costo.

6. Conclusiones y trabajos futuros

En este trabajo de tesis se ha logrado la construcción de un marco comparativo como herramienta para la confrontación de metodologías de minería de datos.

El marco propuesto incluye cuatro aspectos donde se analiza el nivel de especificación de las tareas, los escenarios de aplicación, las actividades que componen cada fase del proceso y las actividades destinadas a la dirección del proyecto. Para cada uno de estos aspectos se propone la evaluación de un conjunto de características que deberían estar presentes en una metodología de minería de datos bien definida.

El marco comparativo ha sido utilizado para confrontar las metodologías CRISP-DM y Catalyst. Aunque ambos enfoques se encuentran actualmente en etapas tempranas de madurez, han logrado cumplir un gran porcentaje de las características evaluadas. Sin embargo, durante el estudio también se han evidenciado los puntos que se deberían mejorar y seguir desarrollando, como las actividades destinadas a la dirección del proyecto.

Si bien en este trabajo se han analizado los cuatro aspectos en función de la proporción de características que se cumplen en cada uno, los resultados podrían ser diferentes si se trabajara con puntajes. Si el usuario del marco comparativo lo considera necesario, puede efectuar dicha valoración según su criterio.

Se espera que el resultado de esta tesis sirva como herramienta para que los equipos de trabajo puedan evaluar metodologías de minería de datos, considerando además la posibilidad de complementar las mismas con los conceptos que no estén presentes o bien utilizar el marco como base para construir una propia.

Como línea de trabajo futuro se podría establecer una ponderación de las características evaluadas mediante un sistema de puntajes que permita obtener resultados cuantitativos respecto al desempeño de cada metodología. El desafío de esta investigación radicará en los criterios que se deberían tener en cuenta para establecer estos valores en forma objetiva. Otra línea de investigación podría ser la ampliación del marco comparativo, incorporando nuevos aspectos al mismo.

Anexo. Técnicas de minería de datos

Las técnicas de minería de datos son algoritmos que tienen por objetivo la extracción de patrones del conjunto de datos.

Las técnicas de clasificación asumen que hay un conjunto de objetos que pertenecen a diferentes clases. La etiqueta de clase es un valor (simbólico) discreto y es conocido para cada objeto. El objetivo es construir modelos de clasificación (a veces llamados clasificadores), que intentarán asignar la etiqueta de clase correcta a nuevos objetos. Los modelos de clasificación son usados principalmente para el modelado predictivo [5].

A continuación se describen las cuatro técnicas de clasificación utilizadas en el caso de estudio de este trabajo de tesis: árboles de decisión, regresión logística, naive bayes y vecinos más próximos.

Arboles de decisión

Los árboles de decisión consisten en una serie de decisiones o condiciones organizadas en forma jerárquica, a modo de árbol, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas [23]. Cuando los árboles de decisión son utilizados para predecir variables categóricas reciben el nombre de árboles de clasificación. En cambio, cuando la variable explicada es continua se construyen árboles de regresión.

En ciertas aplicaciones, especialmente cuando el grupo de predictores contiene una mezcla de variables numéricas y factores, los modelos basados en árboles son más fáciles para interpretar y discutir que los modelos lineales [2].

Por ejemplo, una entidad bancaria desea construir un modelo para determinar qué clientes son potencialmente morosos al momento de pagar un crédito. Se cuenta con información histórica de créditos ya otorgados, donde se conoce el estado civil del cliente, si tiene hijos, el monto del crédito y si resultó moroso para pagar su deuda. Se ha obtenido como resultado el árbol de la Figura 46.

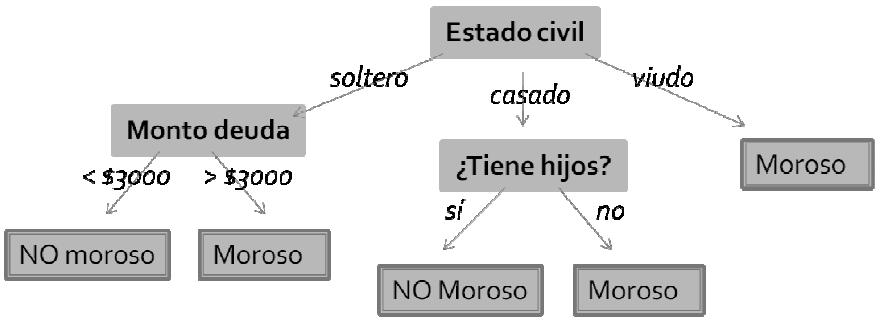

Fig. 46. Árbol de decisión

Cada "nodo" corresponde a un atributo y cada "rama" al valor posible de ese atributo. Una "hoja" del árbol especifica el valor esperado de la decisión de acuerdo con los ejemplos dados. La explicación de una determinada decisión viene dada por la trayectoria desde la raíz a la hoja representativa de esa decisión.

El algoritmo más utilizado es el C4.5 (implementado en WEKA con el nombre J4.8) [36], que utiliza el criterio de ganancia de información (gain ratio) para construir el árbol. Este criterio permite ir construyendo la jerarquía seleccionando el atributo que mejor divide el conjunto de instancias en función de la clase de cada una.

Poda del árbol

Para facilitar la comprensión del árbol puede realizarse una poda del mismo, lo que significa la sustitución de una parte del árbol (sub-árbol) por una hoja. La poda tendrá lugar si el valor esperado de error en el sub-árbol es mayor al de la hoja que lo sustituya.

Vecino más próximo (Nearest neighbors)

En este método cada nuevo caso se compara con los existentes utilizando una métrica de distancia. Se asigna a la nueva instancia la clase mayoritaria entre los casos más próximos (vecinos más cercanos).

Se trabaja con un parámetro *"k"* que determina la cantidad de vecinos cercanos con los cuales se comparará la nueva observación. En su versión más simple k=1, clasificando al nuevo caso con la clase del caso más próximo.

En el ejemplo de la Figura 47 se puede observar cómo cambia el resultado de la clasificación al utilizar diferente valor del parámetro "k". Con k=1 la nueva instancia se clasificará como "rojo", mientras que con k=4 la clasificación será "azul".

Fig. 47. Método vecino más próximo con diferentes valores de k.

El valor del parámetro k debe obtenerse experimentalmente, seleccionando el que genera la menor tasa de error [13].

Una variante del método se basa en determinar una región de cercanía (Fig.48). Cuando un nuevo caso aparece, se genera un circulo con centro en dicho punto y un radio *r* prefijado. Se calcula el número de ejemplos que caen dentro del círculo y se etiqueta al nuevo caso como perteneciente a la clase más numerosa. El valor del radio seleccionado es crítico.

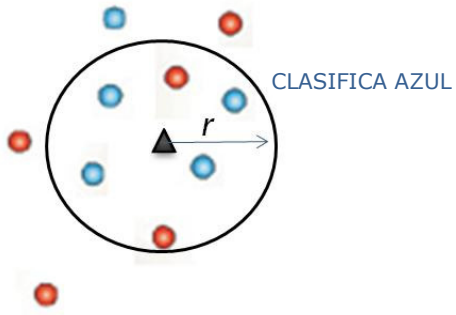

Fig. 48. Método vecino más próximo por región de cercanía.

El método de vecino más próximo no crea un modelo, sino que utiliza todo el conjunto de datos cada vez que hay que clasificar a una nueva instancia. Este tipo de método se denomina "retardado", ya que no crea un modelo general sino que retrasa la decisión de generalización del conjunto de entrenamiento hasta que se presente un nuevo caso, buscando entre los ejemplos almacenados el más parecido y actuando como en esa ocasión.

Clasificador Naive Bayes

Es un clasificador probabilístico basado en el teorema de Bayes. Asume que las variables predictoras (o explicativas) son independientes, conocida la variable de respuesta [32].

Esta técnica puede manejar distintos tipos de variables predictoras tanto numéricas como categóricas.

Dado nuevo caso definido por un conjunto de variables predictoras $\mathbf{X}=\{x_1, x_2, x_n\}$ el método determina la probabilidad de pertenencia a cada una de las clases de la variable de respuesta $\mathbf{C}=\{c_1, c_2, c_m\}$. Esta probabilidad se denomina "a posteriori" y se expresa:

$$P(C_j/x_1,x_2,x_n) = P(x_1,x_2,x_n/C_j).P(C_j) = P(C_j).\prod_{k=1}^{n} P(x_k/C_j)$$

La probabilidad a posteriori indica la probabilidad de que el nuevo caso (caracterizado por el vector de variables predictoras **X**) pertenezca a la clase C_j. La clasificación final se realizará con la clase C_j que tenga la mayor probabilidad a posteriori.

La probabilidad a posteriori de C_j dado X, cuando las variables predictoras son independientes, es el producto de la probabilidad "a priori" de C_j y de la probabilidad de que ocurra **X** cuando ocurrió C_j.

Para ejemplificar el método de Naive Bayes, tomaremos el conjunto de datos de la Tabla 33, donde Y es la variable de respuesta y X_i son las variables predictoras.

X1	X2	X3	Y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
0	0	1	1
1	1	0	1

Tabla 33. Conjunto de datos de ejemplo

Dado un nuevo caso ($X_1 = 0, X_2 = 0, X_3 = 1$) se debe determinar a qué clase pertenece (Y=0/Y=1).

Para clasificar al nuevo caso, se deberán comparar las probabilidades a posteriori:

- $P(Y = 0 / X_1 = 0, X_2 = 0, X_3 = 1) = P(Y = 0).P(X_1 = 0, X_2 = 0, X_3 = 1 / Y = 0)$
- $P(Y = 1 / X_1 = 0, X_2 = 0, X_3 = 1) = P(Y = 1).P(X_1 = 0, X_2 = 0, X_3 = 1 / Y = 1)$

La mayor probabilidad indicará el valor con el que se clasificará a la nueva instancia.

Las probabilidades a priori son:

$P(Y = 0) = 3/7$

$P(Y = 1) = 4/7$

Luego:

$$P(X_1 = 0, X_2 = 0, X_3 = 1 / Y = 0) = P(X_1 = 0 / Y = 0)\, P(X_2 = 0 / Y = 0)\, P(X_3 = 1 / Y = 0) = 2/3 . 1/3 . 1/3 = 2/27$$

$$P(X_1 = 0, X_2 = 0, X_3 = 1 / Y = 1) = P(X_1 = 0 / Y = 1)\, P(X_2 = 0 / Y = 1)\, P(X_3 = 1 / Y = 1) = 2/4 . 2/4 . 3/4 = 3/16$$

Finalmente:

$$P(Y = 0 / X_1 = 0, X_2 = 0, X_3 = 1) = 3/7 . \ 2/27 = 0,03$$

$$P(Y = 1 / X_1 = 0, X_2 = 0, X_3 = 1) = 4/7 . 3/16 = 0,1$$

Como 0,1 > 0,03 el nuevo caso será asignado a la clase 1 (Y=1).

Regresión Logística Binaria

Es un modelo estadístico que permite establecer la relación entre una variable dependiente cualitativa dicotómica o politómica y una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas y predecir la pertenencia a un determinado grupo en términos de probabilidad [6,31].

El modelo de regresión logística binaria es muy utilizado cuando se desea estimar la probabilidad de que ocurra un suceso determinado. Es una herramienta muy flexible en cuanto a las variables explicativas, debido a que permite que las mismas sean numéricas o categóricas.

En esta técnica el problema de clasificación se aborda introduciendo una variable ficticia binaria que representa la ocurrencia o no del suceso. Sea y la variable de respuesta, entonces $y = 1$ si el suceso ocurre o bien $y = 0$ en caso contrario.

En cuanto a los supuestos del modelo, se recomienda que no exista multicolinealidad entre las variables explicativas numéricas, ya que esta situación podría causar importantes sesgos en las estimaciones de la variable dependiente.

Sea **x** el vector formado por las *k* variables explicativas, el primer enfoque para el modelo de regresión sería:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \beta_0 + \boldsymbol{\beta x} + \boldsymbol{\varepsilon}$$

Si llamamos p_i a la probabilidad de que y tome el valor 1 cuando $\mathbf{x}=\mathbf{x_i}$ entonces:

$$p_i = P(y = 1/\boldsymbol{x}_i)$$

La esperanza de y será:

$$E[y/\boldsymbol{x}_i] = P(y = 1/\boldsymbol{x}_i).1 + P(y = 0/\boldsymbol{x}_i).0 = p_i$$

Entonces:

$$p_i = \beta_0 + \boldsymbol{\beta x}_i$$

Por lo tanto, el valor predicho por el modelo de regresión será la probabilidad de que el suceso ocurra ($y = 1$), cuando $\mathbf{x} = \mathbf{x}_i$.

El inconveniente principal de esta formulación es que p_i debería estar acotada en el intervalo [0,1] y no hay ninguna garantía de que la predicción verifique esta restricción. Para garantizar esta situación debemos transformar la variable de respuesta de algún modo tal que:

$$p_i = F(\beta_0 + \boldsymbol{\beta x}_i)$$

De esta forma, garantizaremos que el valor de p_i se encuentre entre cero y uno si exigimos que F tenga esa propiedad.

Habitualmente se toma como F la función de distribución logística, y el modelo lineal resultante se denomina *logit*:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta x_i)}}$$

Los parámetros del modelo son:

- La ordenada al origen β_0
- Los coeficientes de la regresión $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$

El cálculo de los coeficientes del modelo y de sus errores estándar se realiza por estimaciones de máxima verosimilitud, es decir, estimaciones que maximizan la probabilidad de obtener los valores de la variable de respuesta y proporcionados por los datos de la muestra. A diferencia de la regresión lineal múltiple, estas estimaciones no son de cálculo directo, por lo que debemos recurrir a métodos iterativos como el método de Newton Raphson. Como el cálculo es complejo, por lo general se requieren de rutinas de programación o software estadístico.

Para determinar cuánto se modifican las probabilidades por unidad de cambio en las variables explicativas, se utilizan los denominados "odd ratios" o "ratios de probabilidades":

$$O_i = \frac{p_i}{1 - p_i} = e^{(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}$$

Una vez que se han estimado los parámetros del modelo, dado un nuevo caso descripto por el vector de variables explicativas **x**, se puede calcular la probabilidad de que ocurra el suceso (p_i). En general, si p_i > 0.5 entonces se clasifica a la nueva instancia con la clase $y = 1$, caso contrario se clasifica como $y = 0$.

Referencias

1. Azevedo, A., Santos, M.F. (2008). *KDD, SEMMA and CRISP-DM: a parallel overview.* IADIS 2008. Algarve, Portugal.
2. Balzarini, M. G., González, L., Tablada, M., Casanoves, F., Di Rienzo, J. A., & Robledo, C. W. (2008). *INFOSTAT: Manual del Usuario*. Córdoba, Argentina. Editorial Brujas.
3. Britos, P. (2008). *Procesos de explotación de información basados en sistemas inteligentes*. (Tesis doctoral). Universidad Nacional de La Plata. Argentina.
4. Britos, P., Dieste, O., & García-Martínez, R. (2008). *Requirements Elicitation in Data Mining for Business Intelligence Projects*. Advances in Information Systems Research, Education and Practice, 139-150.
5. 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*.
6. Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example* (Vol. 607). Wiley-Interscience.
7. Cios, K. J., & Kurgan, L. A. (2005). *Trends in data mining and knowledge discovery*. Advanced techniques in knowledge discovery and data mining, 1-26.
8. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). *The KDD process for extracting useful knowledge from volumes of data*. Communications of the ACM, 39(11), 27-34.
9. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery in databases*. AI magazine, 17(3), 37.
10. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *Knowledge discovery and data mining: Towards a unifying framework*. Knowledge Discovery and Data Mining, 82-88.
11. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*. MIT Press.
12. Garcia, V. (2010). *Distribuciones de Clases No Balanceadas: Métricas, Análisis de Complejidad y Algoritmos de Aprendizaje*. (Tesis Doctoral). Departament de llenguatges i Sistemes Informàtics, Universitat Jaume I. España.
13. Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques*. Morgan Kaufmann.
14. KDnuggets (2007). *Poll: ¿What main methodology are you using for data mining?* Recuperado el 7 de noviembre de 2010, de http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.

15. Kriegel, H. P., Borgwardt, K. M., Kröger, P., Pryakhin, A., Schubert, M., & Zimek, A. (2007). *Future trends in data mining*. Data Mining and Knowledge Discovery, 15(1), 87-97.

16. Kurgan, L. A., & Musilek, P. (2006). *A survey of Knowledge Discovery and Data Mining process models.* Knowledge Engineering Review, 21(1), 1-24.

17. Marbán, Ó., Mariscal, G., Menasalvas, E., & Segovia, J. (2007). *An engineering approach to data mining projects*. Intelligent Data Engineering and Automated Learning-IDEAL 2007, 578-588.

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

19. Mariscal, G., Marbán, Ó., & Fernández, C. (2010). *A survey of data mining and knowledge discovery process models and methodologies.* Knowledge Engineering Review, 25(2), 137.

20. Mariscal, G., Marbán, Ó., González, Á. L., & Segovia, J. (2007). *Hacia la Ingeniería de Data Mining: Un modelo de proceso para el desarrollo de proyectos*. II Congreso Español de Informática, V Taller de Minería de Datos y Aprendizaje (TAMIDA '07). Zaragoza, España.

21. Moyle, S., & Jorge, A. (2001). *RAMSYS - A methodology for supporting rapid remote collaborative data mining projects*. ECML/PKDD01, Workshop Integrating Aspects of Data Mining, Decision Support and Meta-learning. Freiburgo, Alemania.

22. Nascimento, G., & Oliveira, A. (2012). *AgileKDD: An Agile Knowledge Discovery in Databases Process Model*. II International Conference on Advances in Information Mining and Management (IMMM 2012). Venecia, Italia.

23. Orallo, J. H., Quintana, M. J. R., & Ramírez, C. F. (2004). *Introducción a la Minería de Datos.* Pearson Prentice Hall.

24. Pressman, R. (2005). *Ingeniería de software: un enfoque práctico.* McGraw-Hill.

25. Project Management Institute (2005). *Guía de los fundamentos de la dirección de proyectos (Guía del PMBOK®).* Tercera edición.

26. Pyle, D. (2003). *Business Modeling and Data Mining*. Morgan Kaufmann Publishers.

27. Rodríguez, D., Pollo-Cattaneo, F., Britos, P., & García Martínez, R. (2010). *Estimación empírica de carga de trabajo en proyectos de explotación de información*. XVI Congreso Argentino de Ciencias de la Computación (CACIC 2010). Buenos Aires, Argentina.

28. R-project. *The R Project for Statistical Computing*. Recuperado el 25 de octubre de 2011, de http://www.r-project.org.
29. SAS Institute (1998). *Data Mining and the Case for Sampling*. Recuperado el 14 de abril de 2012, de http://nas.uhcl.edu /boetticher/ML_DataMining/SAS-SEMMA.pdf.
30. SAS Institute. *SEMMA*. Recuperado el 10 marzo de 2011, de http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html.
31. Sociedad Española de Bioquímica Clínica y Patología Molecular. *Regresión Logística*. Recuperado el 2 de mayo de 2012, de http://www.seqc.es/es/Varios/7/40/Modulo_3:_Regresion_logistica_y_ multiple.
32. StatSoft, Inc. (2012). *Electronic Statistics Textbook*. Recuperado el 15 de noviembre de 2012, de http://www.statsoft.com/textbook.
33. TuTiempo Network. *Clima Mundial, Datos Históricos Climáticos*. Recuperado el 02 de septiembre de 2011, de http://www.tutiempo.net /clima.
34. University of Waikato. *WEKA, Data Mining Software in JAVA*. Recuperado el 15 de agosto de 2011, de http://www.cs.waikato.ac.nz /ml/weka.
35. Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining*. IV International Conference on the Practical Applications of Knowledge Discovery and Data Mining (pp. 29-39).
36. Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.
37. Yang, Q., & Wu, X. (2006). *10 challenging problems in data mining research*. International Journal of Information Technology & Decision Making, 5(04), 597-604.