Metodologías y estándares

Jordi Gironés Roig

PID_00197285

Universitat Oberta de Catalunya

Índice

Introducción

Al finalizar este módulo el estudiante comprenderá la necesidad y utilidad de una metodología para la gestión de proyectos de minería de datos, de un estándar de comunicación de resultados de modelos, de un conjunto de buenas prácticas para el gobierno de servicios IT, y de una visión estratégica y de negocio de las actividades analíticas en toda organización.

La calidad como fin último en la consecución de objetivos es el concepto que sustenta la necesidad que la industria tiene de utilizar metodologías, tanto para llevar a cabo proyectos de puesta en marcha de nuevas funcionalidades como para ofrecer servicios de funcionalidades existentes.

Solo si hay una planificación, un seguimiento, una ejecución, una revisión y una verificación podemos garantizar una calidad en el producto que finalmente se elabora. El estudiante aprenderá a adaptar la implementación de la calidad en entornos relacionados con BA.

1. Metodologías y estándares

Nos enfrentamos a proyectos complejos con multitud de tareas interdisciplinares e interdependientes, que además mezclan intereses y necesidades de diferentes grupos de personas y que normalmente están condicionados por limitaciones económicas y tecnológicas.

Lo recomendable en estos casos es diseñar una hoja de ruta que nos va a permitir saber dónde estamos, dónde queremos llegar y las medidas a tomar para corregir periódicamente las desviaciones del rumbo seguido.

Las hojas de ruta que proponemos son:

- Metodología CRISP-DM para la gestión de proyectos de minería de datos.
- Factores delta, factores clave para cultivar la visión analítica en las organizaciones.
- PMML como lenguaje estándar de apoyo al despliegue y mantenimiento de modelos *data mining*.
- Norma ISO 20000 para el gobierno de servicios informáticos.

Se trata sin duda de cuatro guías que contribuirán enormemente a la consecución de objetivos, a la mejora de procesos y a la adopción de una cultura empresarial soportada por el análisis y el estudio de la información.

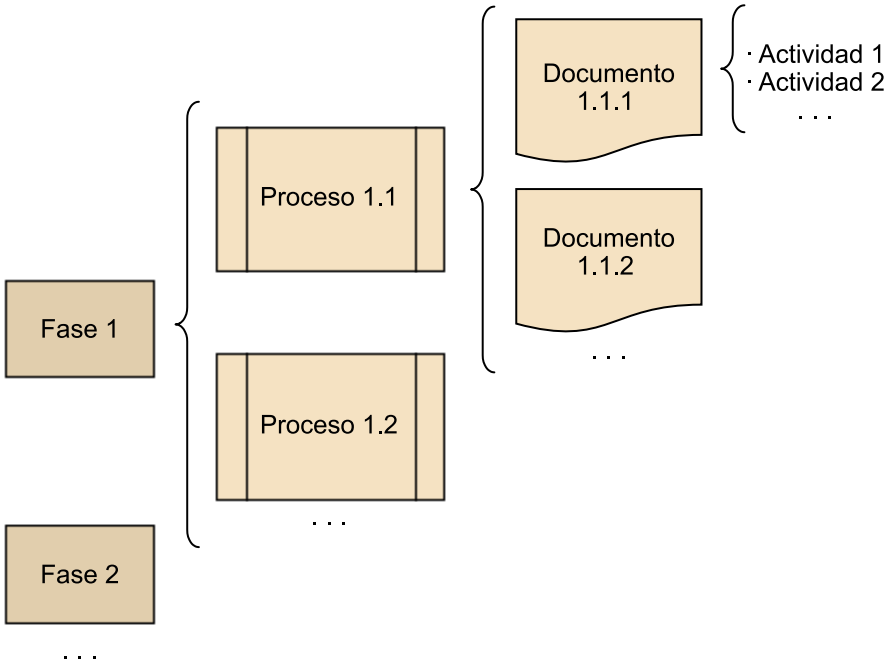
1.1. Metodología CRISP-DM

CRISP-DM (*cross industry standard process for data mining*) nació en el seno de dos empresas, DaimlerChrysler y SPSS, que en su día fueron pioneras en la aplicación de técnicas *data mining* en los procesos de negocio. CRISP-DM es una metodología basada en la práctica y experiencia real de analistas DM que han contribuido activamente al desarrollo de la misma.

CRISP-DM se organiza en fases, procesos, documentos entregables y actividades.

Una fase, por ejemplo "Comprensión del negocio" se subdivide en procesos como "Determinar los objetivos del negocio" e "Identificar los objetivos *data mining*". A su vez, los procesos tienen documentos entregables a los que llegaremos después de haber ejecutado una lista de actividades.

Figura 1. Estructura de la metodología (CRISP-DM 1.0)



Calidad total

Un aspecto a destacar es que la iteración y revisión de fases y procesos se establece como un aspecto clave si se quiere ejecutar un proyecto de calidad.

De este modo se establecen micro ciclos de planificación, ejecución y revisión, de los que solo se sale cuando el proceso de revisión es satisfactorio. Este principio está muy presente tanto en la norma ISO 9000 como en la ISO 20000.

Todas las fases son importantes, por supuesto, pero quisiera remarcar que la tendencia natural de la condición humana, por experiencia propia, es la de concentrar recursos en exceso al final del proyecto, en la fase despliegue, por no haber hecho las cosas bien en las fases anteriores.

Merece la pena y es más óptimo y económico, no escatimar recursos en las fases iniciales de preparación, planificación, construcción e iteración.

Vamos a estudiar con detalle todas las fases que nos propone la metodología CRISP-DM. Observad que en el centro del esquema que la resume se encuentra el objetivo de la misma, es decir, la conversión de los datos en conocimiento.

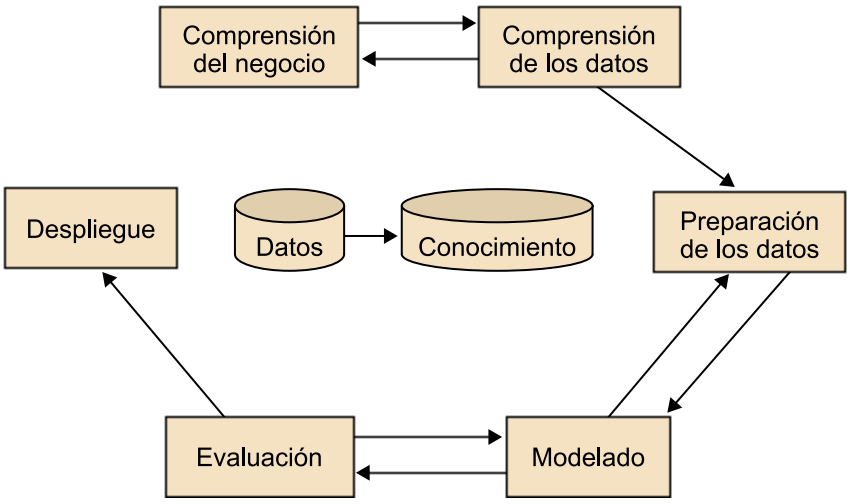
Stakeholder

A lo largo del estudio de las fases hablaremos repetidamente de "el cliente" al que hay que satisfacer. Sería muy útil hacer una generalización de este concepto, en inglés *stakeholder* o "partes interesadas", no confundir con los usuarios clave.

Parte interesada es todo aquel a quien debemos reportarle directa o indirectamente cuentas del proyecto, precisamente por el interés que tiene en el mismo. A partir de ahora, cliente será equivalente a parte interesada.

La siguiente figura esquematiza el ciclo de fases que propone CRISP-DM.

Figura 2. Fases de la metodología CRISP-DM (CRISP-DM 1.0)



Adecuación de la metodología al proyecto

Merece la pena mencionar que la metodología debe ser entendida siempre como una guía de trabajo que permite garantizar una calidad en la entrega del proyecto. Para conseguir que efectivamente sea una guía de trabajo útil y práctica, deberemos adaptarla a las necesidades de nuestro proyecto en concreto.

Por la propia idiosincrasia de nuestro proyecto merecerá más la pena desarrollar a fondo la metodología en algunas fases; y quizá en otras, o porque muchas cosas nos vienen dadas o porque ya se han ejecutado con éxito en proyectos anteriores, no merecerá la pena ser tan exhaustivos en la documentación, control y gestión de las mismas.

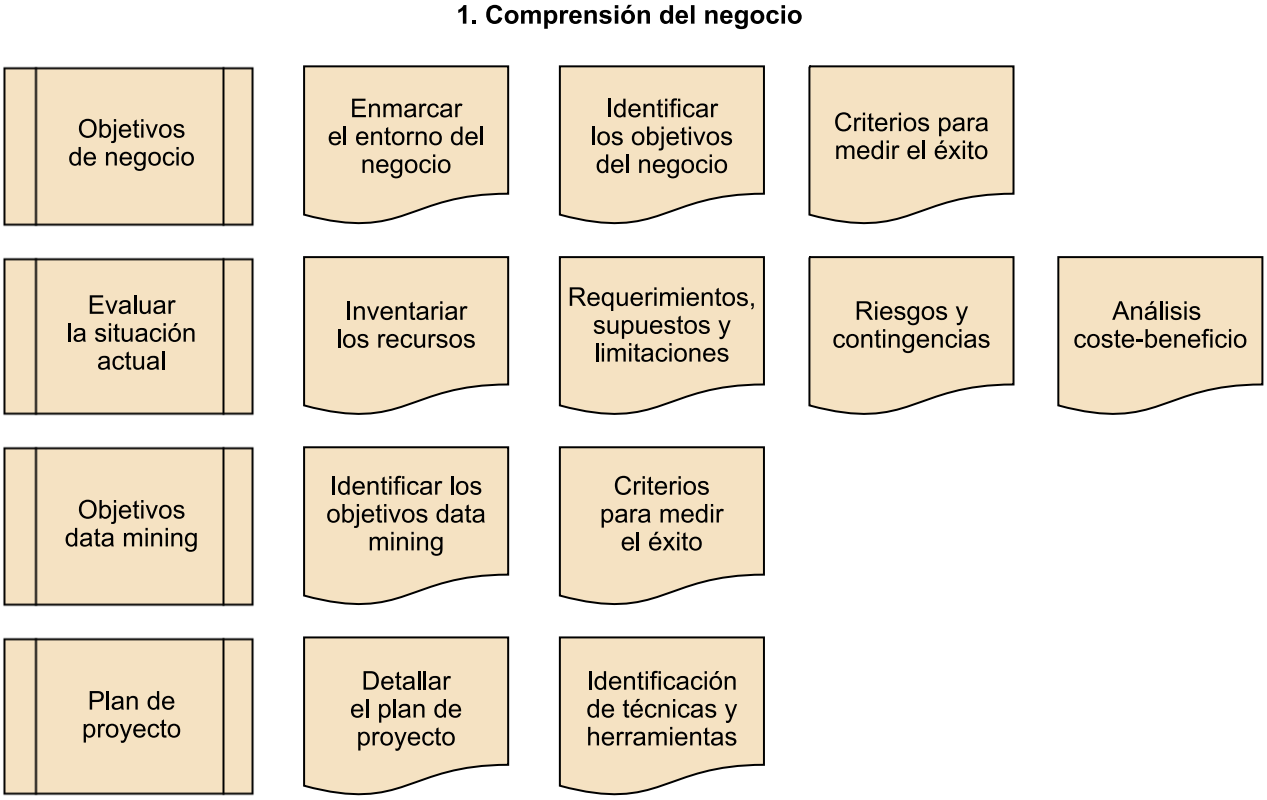
1.1.1. Comprensión del negocio

En esta fase trataremos de conseguir desde una clara perspectiva de negocio cuáles son los objetivos del mismo, tratando de evitar el gran error de dedicar el esfuerzo de todo el proyecto a proporcionar respuestas correctas a preguntas equivocadas.

Con los objetivos de negocio en mente, elaboraremos un estudio de la situación actual del negocio respecto de los objetivos planteados, en este punto, trataremos de clarificar recursos, requerimientos y limitaciones, para así poder concretar objetivos *data mining* que contribuyan claramente a la consecución de los objetivos primarios.

Finalmente, elaboraremos un plan de proyecto en el que detallaremos las fases, tareas y actividades que nos deberán llevar a alcanzar los objetivos planteados.

Figura 3. Comprensión del negocio (CRISP-DM 1.0)



Objetivos del negocio

Enmarcar el entorno del entorno

Recogeremos en un documento la situación actual de la organización, tratando de este modo de establecer una fotografía del punto de partida del proyecto. En el documento debe aparecer una primera aproximación de los objetivos de negocio, así como de los recursos tanto materiales como humanos con los que se cuenta para ello.

Actividades a ejecutar son:

- Identificar en la organización, divisiones, departamentos y grupos de trabajo.
- Identificar personas clave en la organización, sus funciones y responsabilidades.
- Identificar al patrocinador del proyecto, los *stakeholders* del proyecto y proponer un comité de seguimiento para el mismo.
- Identificar las unidades de negocio implicadas en el proyecto.

- Describir el problema que tratamos de resolver con el proyecto e identificar los antecedentes del mismo (ya se ha hecho algo al respecto, se llevan a cabo tareas analíticas en algún departamento, etc.).

- Identificar a los usuarios clave y documentar sus necesidades y expectativas en el proyecto.

- Esclarecer si ya se han tomado iniciativas para cubrir las necesidades detectadas y en caso afirmativo, valorar ventajas e inconvenientes de estas iniciativas.

Identificar los objetivos del negocio

Identificar el objetivo principal de la organización sería como el *leitmotiv* del proyecto, al que le debería seguir una lista de objetivos secundarios que ayudarán a concretar y contribuir a alcanzar el objetivo principal. Es importante remarcar que los objetivos deben ser alcanzables.

Actividades a ejecutar son:

- Descripción informal del problema que se pretende resolver.

- Listar las preguntas de negocio a las que se pretende dar respuesta con el proyecto.

- Identificar requerimientos colaterales en el proyecto.

- Listar los beneficios que se pretenden conseguir con el proyecto.

Criterios para medir el éxito

Cada objetivo de negocio debe poder ser asociado al menos a un criterio medible de éxito y a ser posible, habrá de establecer quién ejecutará estas mediciones.

Medir el grado de cumplimiento de los objetivos de negocio es un aspecto irrenunciable del proyecto y debe ser tenido en cuenta en el momento de plantear los propios objetivos.

Evaluar la situación actual

Detallaremos los recursos a nivel de hardware, software, fuentes de datos y personal de que disponemos, también detallaremos los requerimientos, supuestos que aceptamos y limitaciones identificadas, para, finalmente, diseñar una matriz de riesgos y contingencias.

Inventario de recursos

A nivel de recursos humanos identificaremos a expertos de negocio, soporte tecnológico y expertos analistas.

A nivel de datos, identificaremos fuentes de datos externas, internas, analíticas y operacionales.

A nivel de recursos de computación, identificar hardware y estado en el que se encuentra. Identificar también recursos de software *data mining* o analítico disponible o a adquirir.

Requerimientos, supuestos y limitaciones

Estableceremos los requerimientos respecto de la programación del proyecto, precisión, capacidad de despliegue, capacidad de mantener los servicios puestos en marcha y capacidad de repetir los pasos de modelado y despliegue para futuros ajustes.

Asimismo, estableceremos los supuestos en términos de objetivos marcados, calidad de los datos, factores externos como coyuntura económica o acciones de la competencia. Clarificar más las estimaciones económicas que se hayan hecho para adquirir recursos. Clarificar también los supuestos acordados en cuanto a comunicación y explicación del modelado utilizado a la dirección.

Identificar las limitaciones en diferentes ámbitos como el legal, el presupuestario o los recursos de todo tipo. Por ejemplo, derechos legales de acceso a los datos, accesibilidad tecnológica a los datos, presupuesto de costes fijos, costes de implementación y rangos de tolerancia en las desviaciones.

Riesgos y contingencias

Deberemos identificar los riesgos del proyecto, que pueden venir por problemas en el negocio, en la propia organización interna, en los recursos económicos, en los aspectos tecnológicos o en la baja calidad de las fuentes de datos. Considerar también circunstancias que podrían impactar en el proyecto, así como su coste en tiempo y en dinero.

Para mitigar o anular los riesgos deberemos prever un plan de contingencias.

Análisis coste-beneficio

Deberemos estimar los costes de adquisición de datos, costes de implementación y despliegue de la solución, costes de formación, partida para imprevistos, costes operativos de mantenerla en funcionamiento y finalmente, un estudio de los beneficios esperados, acompañado de un estudio ROI del retorno de la inversión esperados.

Objetivos *data mining*

Si los objetivos de negocio están expuestos en términos empresariales, los objetivos de *data mining* deben estar expuestos en términos técnicos dentro del ámbito de conocimiento de la minería de datos.

Identificar los objetivos *data mining*

Deberemos hacer el ejercicio de trasladar los objetivos de negocio a la arena del *data mining*. También será conveniente asignar cada objetivo a la correspondiente competencia *data mining*, es decir, clasificación, asociación, segmentación, predicción, etc.

Criterios para medir el éxito

Estableceremos los grados de precisión que exigiremos a nuestros modelos para ser aceptados, estos criterios deberán estar soportados por las buenas prácticas del mercado (*benchmarking*).

Plan de proyecto

Merece la pena mencionar que el plan de proyecto debe concebirse como una herramienta dinámica y susceptible de ser revisada, actualizada y modificada siempre que sea necesario y que debería ser consultado si no al inicio y finalización de cada tarea, sí al menos en los hitos.

Detallar el plan de proyecto

Listaremos las fases identificadas junto con su respectivo detalle de tareas, duración, recursos necesarios, entradas y salidas de información, y dependencias. Importante también marcar los riesgos identificados y especificar el impacto en tiempo que podrían acarrear al proyecto.

Será de gran ayuda poder plasmar en el plan de proyecto la iteración de fases hasta conseguir los niveles de calidad establecidos e identificar los puntos de decisión y revisión que se establezcan.

Evaluación inicial de herramientas y técnicas

Estableceremos los criterios de selección para la herramienta *data mining* que se vaya a seleccionar. Elaboraremos una lista de posibles software que cumplan con los criterios establecidos.

Se evaluarán también la oportunidad de uso de determinadas técnicas *data mining* teniendo en cuenta las necesidades del proyecto y las capacidades de la herramienta seleccionada.

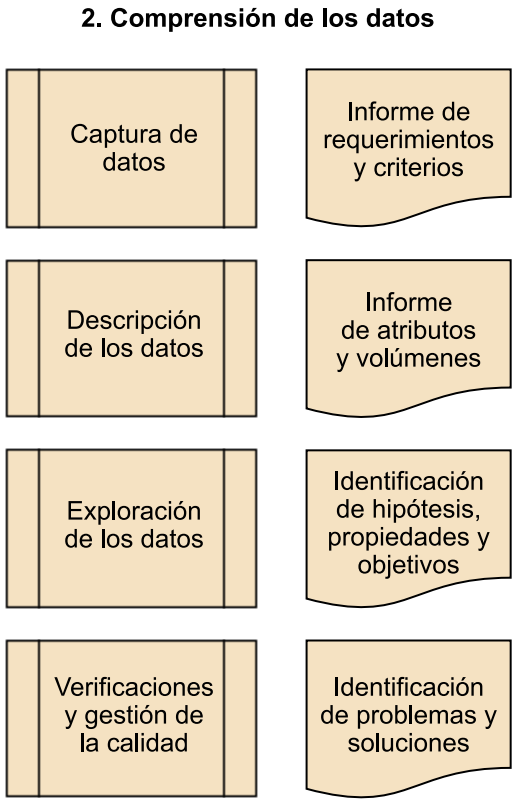
1.1.2. Comprensión de los datos

Comprensión se refiere a trabajar los datos con el objetivo de familiarizarse al máximo con ellos, saber de dónde provienen, en qué condiciones nos llegan, cuál es su estructura, qué propiedades tienen, qué inconvenientes presentan y cómo podemos mitigarlos o eliminarlos.

Se trata de una fase crítica puesto que es donde trabajamos de lleno con la calidad de los datos, que por otro lado debemos ver como la materia prima para el *data mining*.

Tener una buena calidad de los datos será siempre una condición necesaria aunque no suficiente para tener éxito en el proyecto.

Figura 4. Comprensión de los datos (CRISP-DM 1.0)



Captura de datos

Ejecutaremos los procesos de carga de información con el objetivo de iniciar las tareas de comprensión de los datos. Existen herramientas especializadas en los procesos de comprensión de los datos o bien la propia herramienta *data mining* puede llegar a cubrir las necesidades de esta fase.

Informe de requerimientos y criterios

Documentaremos las distintas selecciones realizadas a la hora de cargar información, así como si hay atributos más importantes que otros.

Verificaremos si disponemos de los datos y atributos necesarios para alcanzar nuestros objetivos. Prestaremos especial atención a los procesos de integración de información de varias fuentes en una, puesto que pueden generar problemas nuevos.

Se deberá tener prevista la gestión de los valores ausentes *missing values* o incluso la gestión de datos no en formato electrónico, en papel u otros.

Descripción de los datos

En esta fase realizaremos los primeros pasos de exploración de los datos.

Informe de atributos y volúmenes

Por un lado documentaremos tanto el formato de los datos que nos llegan como su nivel de calidad, inventariaremos las tablas con las que trabajaremos, sus relaciones y volumetría.

Por otro lado, analizaremos posibles correlaciones entre atributos, así como calcular medidas básicas de estadística, como la media, mediana, desviación estándar, variancia, moda, etc.

Es importante que tratemos de relacionar estas medidas básicas con el negocio, intentando encontrar explicaciones de por qué obtenemos estos datos calculados.

Deberemos asegurarnos de si todos los atributos de trabajo tienen o no alguna relación con los objetivos planteados y verificaremos con la ayuda de expertos de negocio si todavía deberían incorporarse más atributos de trabajo.

Quizá de este informe salga la necesidad de revisar los documentos tanto de objetivos como de supuestos.

Exploración de los datos

La utilización de las técnicas clásicas de exploración de datos, *queries*, informes y gráficos pueden ayudarnos a confirmar o revisar los objetivos *data mining* planteados, a revisar los procesos de captura y descripción de datos y averiguar si hay que incorporar tareas de transformación de datos.

Identificación de hipótesis, propiedades y objetivos

Documentaremos las conclusiones de esta primera exploración de datos, entre las que incluiremos planteamientos de hipótesis iniciales y su conversión en objetivos *data mining* si es posible, contribuiremos a clarificar y concretar más si cabe los objetivos *data mining*.

En cierto modo, la exploración que se propone consiste en una búsqueda a ciegas porque no sabemos ni qué buscamos ni qué nos encontraremos. Dicho esto, será preferible que esta búsqueda siempre se haga con perspectiva de cubrir los objetivos *data mining* marcados.

Verificaciones y gestión de la calidad

En esta fase se desarrollarán la mayor parte de las actividades *data quality management*, que se tratarán más extensamente en un capítulo propio.

Identificación de problemas y soluciones

- Estudiar el grado de cobertura de los datos ¿están todos los casos posibles representados? o si por el contrario, lo que tenemos en realidad es una visión sesgada del universo que queremos estudiar.

- Buscaremos inconsistencias en los datos, precios desorbitados, ingresos imposibles, etc.

- Identificaremos atributos vacíos y establecer una estrategia para reemplazarlos o eliminarlos.

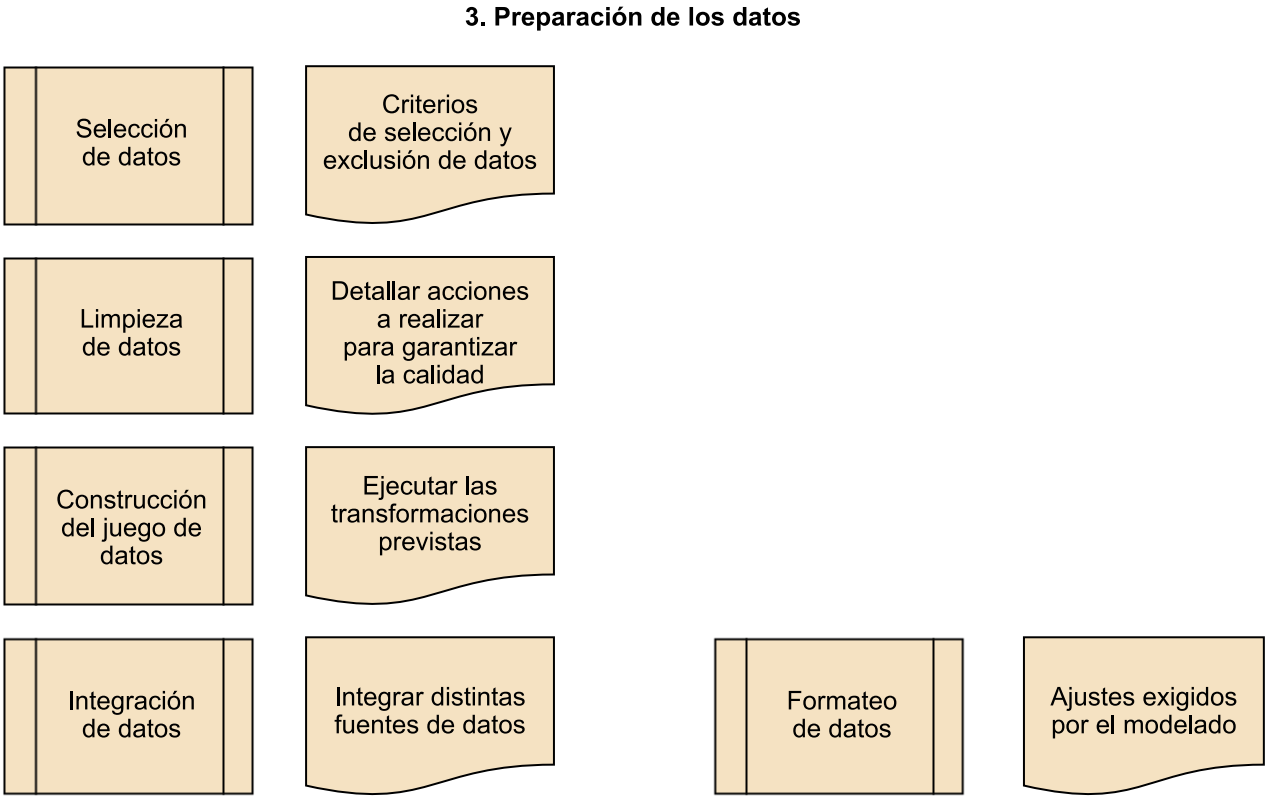
- Estudiar las desviaciones por si se trata de ruido, valores *outliers*, por ejemplo, o si por el contrario se trata de patrones que merecen más estudio.

- Contrastar los supuestos hechos con anterioridad para verificar si una vez revisados los datos, siguen teniendo sentido o hay que replantearlos.

1.1.3. Preparar los datos

El objetivo de esta fase es el de poder disponer del juego de datos final sobre el que se aplicarán los modelos. También se desarrollará la documentación descriptiva necesaria sobre el juego de datos.

Figura 5. Preparación de los datos (CRISP-DM 1.0)



Selección de datos

Deberemos dar respuesta a la pregunta ¿qué datos son los más apropiados para alcanzar los objetivos marcados? Esto significa evaluar la relevancia de los datos, la calidad de los mismos y las limitaciones técnicas que se puedan derivar de aspectos como el volumen de datos.

Documentaremos los motivos tanto para incluir datos, como para excluir datos.

Criterios de selección y exclusión de datos

Nos replantearemos los criterios de selección de datos basándonos, por un lado, en la experiencia adquirida en el proceso de exploración de datos, y por otro lado, en la experiencia adquirida en el proceso de modelado.

Consideraremos el uso de técnicas estadísticas de muestreo y técnicas de relevancia de atributos, que nos ayudarán, por ejemplo, a plantear la necesidad de iniciar actividades de reducción de la dimensionalidad.

Prestaremos atención a la incorporación de datos de diferentes fuentes y por supuesto a la gestión del ruido.

Limpieza de datos

En este paso ejecutaremos tareas derivadas de la gestión de la calidad de los datos, como la gestión de datos ausentes vía el relleno con valores por defecto o mediante técnicas estadísticas como la estimación de valores.

Detallar acciones a realizar para garantizar la calidad

Gestionaremos el ruido, ignorándolo y documentándolo, eliminándolo o corrigiéndolo.

Gestionaremos aquellos valores especiales que pueden llevar a conclusiones erróneas, por ejemplo, preguntas no respondidas en cuestionarios, o truncados de valores numéricos.

Construcción del juego de datos

Se ejecutarán las tareas propias de construcción del juego de datos, extracción de datos de acuerdo a los criterios de selección establecidos, generación de nuevos atributos calculados, transformación de atributos existentes.

Ejecutar las transformaciones previstas

Nos plantearemos el uso de herramientas para ejecutar estas tareas en función de si vamos a conseguir más eficiencia, precisión y capacidad de repetir las tareas de construcción del juego de datos. Será deseable tener la posibilidad de construir juegos de datos por escenarios o simulaciones.

La generación de nuevos atributos puede venir motivada por una necesidad de normalización del atributo, por una consecuencia del modelado que saque a la luz aspectos no cubiertos con los datos existentes o por ejemplo, porque nos interesa, por el tipo de algoritmo, convertir un atributo numérico en categórico.

En esta fase también deberemos considerar las necesidades del algoritmo seleccionado, por ejemplo, si vamos a usar una regresión lineal, deberemos considerar si hay atributos con una relación no lineal respecto de la variable objetivo, puesto que estos atributos no deberían ser usados en la fase de modelado.

Integración de datos

Ejecutaremos tareas de generación de nuevos registros procedentes de tablas o fuentes distintas, se tratará de gestionar las complejidades propias de las fusiones de datos.

Integrar las distintas fuentes de datos

La agregación suele ser una de las actividades que se lleva a cabo en un proceso de fusión de datos. Pasamos de un estado de información detallada a un estado de información sumarizada.

Formateo de datos

El formateo se refiere a cambios sobre los atributos, que solo modifican su forma y nunca su significado.

Ajustes exigidos por el modelado

Habrá que reorganizar atributos, por ejemplo, los atributos clave al principio y el atributo objetivo al final.

En función de las necesidades del modelado, podría ser conveniente ordenar los registros de un determinado modo.

Finalmente, se ejecutarán tareas de formateo de los propios atributos.

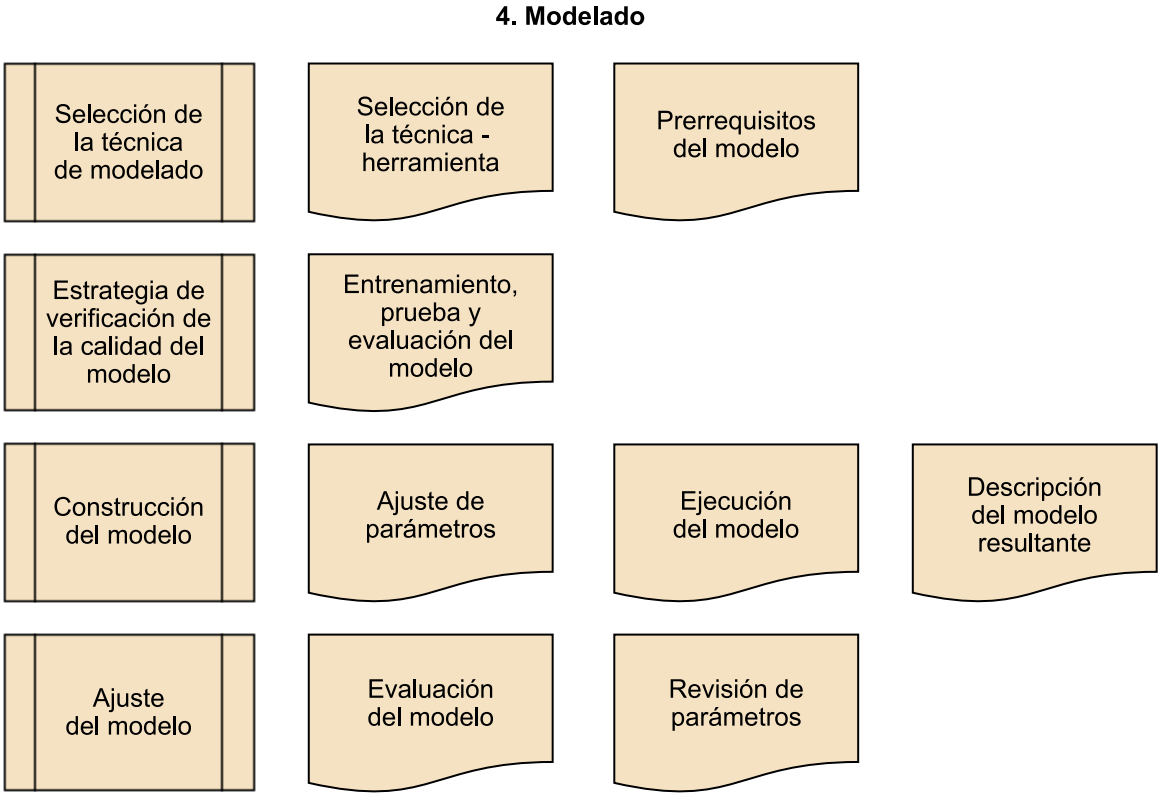
1.1.4. Modelado

El objetivo último de esta fase será el de disponer de un modelo que nos ayude a alcanzar los objetivos *data mining* y los objetivos de negocio establecidos en el proyecto.

Podemos entender el modelo como la habilidad de aplicar una técnica a un juego de datos con el objetivo de predecir una variable objetivo o encontrar un patrón desconocido.

El hecho de que esta fase entre en iteración tanto con su antecesora, la preparación de los datos, como con su sucesora, la evaluación del modelo, nos da una idea de la importancia de la misma en términos de la calidad del proyecto.

Figura 6. Modelado (CRISP-DM 1.0)



Selección de la técnica de modelado

Dado un problema en el ámbito *data mining*, pueden existir una o varias técnicas que den respuesta al mismo, por ejemplo:

- Un problema de segmentación puede aceptar técnicas de *clustering*, de redes neuronales o simplemente técnicas de visualización.

- Un problema de clasificación puede aceptar técnicas de análisis discriminante, de árboles de decisión, de redes neuronales o de K Nearest Neighbor.

- Un problema de predicción aceptará técnicas de análisis de regresión, de árboles de regresión, de redes neuronales o de K-NN.

- Un problema de análisis de dependencias puede afrontarse con técnicas de análisis de correlaciones, análisis de regresión, reglas de asociación, redes bayesianas o técnicas de visualización.

En definitiva, un mismo problema puede resolverse con varias técnicas y una técnica puede servir para resolver varios problemas.

Selección de técnicas y herramientas

Del universo de posibilidades deberá seleccionarse una o varias técnicas para ser usadas por separado, varios modelos, o en combinación, generando un único modelo en varios pasos.

Identificar los prerrequisitos del modelo

Los datos deben estar en formatos específicos, los atributos en posiciones concretas, los registros en un orden preestablecido, las relaciones entre atributos quizá deben cumplir condiciones de independencia o de linealidad.

Deberemos verificar todos los requisitos que nos exija la técnica seleccionada y regresar a la tarea de preparación de los datos en caso necesario.

Estrategia de verificación de la calidad del modelo

Deberemos diseñar el procedimiento a seguir para verificar el grado de precisión del modelo o para entrenarlo si es necesario. Por ejemplo, en el caso de técnicas supervisadas como la clasificación, el juego de datos deberá separarse en subjuegos para el entrenamiento del modelo, las pruebas del modelo y la verificación del modelo, cumpliendo unos parámetros de precisión preestablecidos.

Entrenamiento, prueba y evaluación del modelo

Será deseable preparar un juego de entrenamiento, prueba y verificación específicos para cada objetivo *data mining*.

En el caso de requerir iteraciones en el proceso de construcción, deberemos decidir o calcular cuántas, teniendo en cuenta los efectos negativos del sobreentrenamiento y de la pérdida de tiempo y recursos computacionales.

Construcción del modelo

Se ejecutará la herramienta de modelado sobre los juegos de datos con el fin de crear uno o varios modelos.

Ajustes de parámetros

Los algoritmos suelen tener parámetros que deberemos ajustar y documentar para dejar constancia de la base lógica que los justifica.

Ejecución del modelo

Ejecutaremos la técnica seleccionada sobre nuestro juego de datos y se dejará constancia de los resultados obtenidos.

Descripción del modelo resultante

Documentaremos tanto los parámetros utilizados como el resultado de aplicar el modelo sobre el juego de datos. Por ejemplo, en el caso de modelos basados en reglas, documentaremos las reglas obtenidas así como su grado de cobertura del problema y precisión en el resultado.

En el caso de modelos opacos, como las redes neuronales, documentaremos la topología de la red, el grado de precisión de la misma o el grado de sensibilidad observada.

Prestaremos especial atención a las consecuencias que pudieran derivarse de los resultados, como la identificación de patrones.

Ajuste del modelo

Anteriormente se definieron los criterios de éxito para los objetivos *data mining*, pues bien, en esta tarea procederemos a la verificación de su cumplimiento.

Evaluación del modelo

Será recomendable crear un ranking de los resultados obtenidos en función del grado de cumplimiento de los criterios de éxito sobre los objetivos *data mining*.

Interpretaremos los resultados en clave de negocio, contando para ello con expertos de negocio y por supuesto con el analista.

Evaluaremos la plausibilidad y fiabilidad del modelo así como su impacto en los objetivos *data mining* establecido. ¿Sugieren estos resultados el planteamiento de nuevos objetivos?

Contrastaremos los resultados contra nuestra base de datos de conocimiento para verificar si las conclusiones aportan conocimiento nuevo y útil.

Analizaremos opciones y potencialidades de aplicación y despliegue de los resultados.

Si la salida del modelo consiste en la generación de reglas, nos aseguraremos de que tienen sentido, son aplicables y que el número de reglas generado es el adecuado.

Revisión de parámetros

Analizaremos el impacto que los parámetros han tenido en el resultado final, con el objetivo de ganar conocimiento en el proceso de ajuste del modelo y poder iterar los pasos de modelado y evaluación del modelo hasta dar con el mejor modelo posible.

1.1.5. Evaluación del modelo

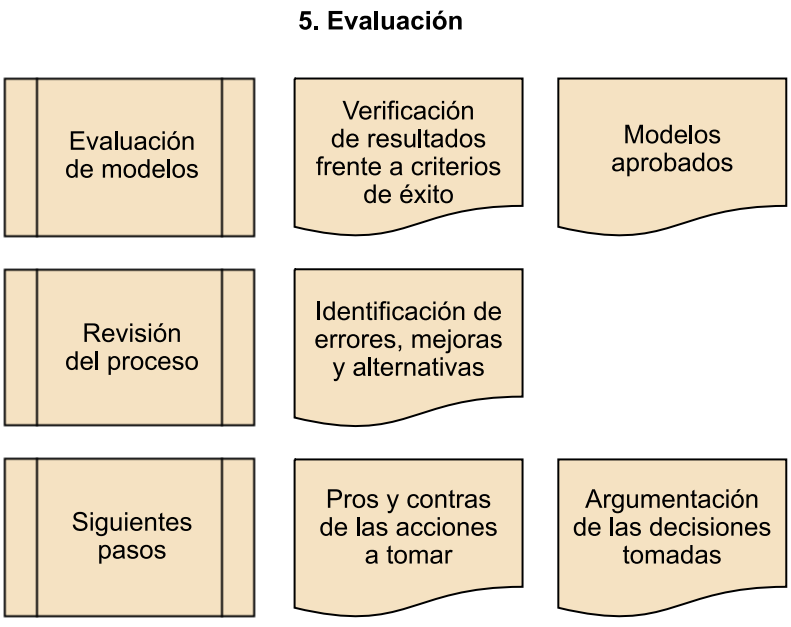
En fases anteriores nos hemos preocupado de asegurar la fiabilidad y plausibilidad del modelo, en cambio en esta fase nos centraremos en evaluar el grado de acercamiento a los objetivos de negocio y en la búsqueda, si las hay, de razones de negocio por las cuales el modelo es ineficiente.

Una forma esquemática y gráfica de visualizar el propósito de un proyecto *data mining* es pensar en la siguiente ecuación:

$$Resultados = Modelos + Descubrimientos$$

Es decir, el propósito de un proyecto *data mining* no son los modelos, que son por supuesto importantes, sino también los descubrimientos, que podríamos definir como cualquier cosa aparte del modelo que contribuye a alcanzar los objetivos de negocio o que contribuye a plantear nuevas preguntas, que a su vez son decisivas para alcanzar los objetivos de negocio.

Figura 7. Evaluación (CRISP-DM 1.0)



Evaluación de modelos

Siempre y cuando sea posible probaremos el modelo en entornos de prueba para asegurarnos de que el posterior proceso de despliegue se realiza satisfactoriamente y para asegurarnos también de que el modelo obtenido es capaz de dar respuesta a los objetivos de negocio.

Verificación de resultados contra criterios de éxito

Documentaremos los resultados de nuestras evaluaciones en términos de cumplimiento de los objetivos de negocio, cuantificándolos si es posible y contrastándolos con los criterios de éxito establecidos.

Estableceremos un ranking de resultados con respecto a los criterios de éxito con relación al grado de cumplimiento de los objetivos de negocio.

Adicionalmente, también se emitirá opinión sobre otros descubrimientos que se hayan realizado aparte del modelado, que aunque probablemente no contribuyan directamente a los objetivos planteados, quizá puedan abrir puertas a nuevos planteamientos y líneas de trabajo.

Modelos aprobados

Argumentaremos la decisión de aprobación o no de los modelos, haciendo referencia a los resultados y a los criterios de éxito establecidos.

Revisión del proceso

En este punto disponemos de uno o varios modelos aprobados y que en principio cumplen con los objetivos de negocio planteados. Se impone ahora una revisión genérica de todo el proceso de minería de datos para asegurar que no se ha tratado algún aspecto importante de forma demasiado superficial.

La motivación principal de esta tarea de revisión será la de realizar una revisión integral de los niveles de calidad del proyecto.

Identificación de errores, mejoras y alternativas

Identificaremos tareas que probablemente se hayan ejecutado sin suficiente rigor y merezcan ser revisadas o repetidas.

Trataremos de identificar mejoras y alternativas de optimización sobre tareas y actividades.

Siguientes pasos

Como consecuencia de todas las tareas de evaluación ejecutadas en esta fase, en esta tarea tomaremos la decisión de repetir o revisar algunos pasos o incluso fases enteras, o bien damos por finalizada la fase de construcción y pasamos al despliegue, o incluso fruto del conocimiento adquirido hasta ahora, podría en este punto proponerse el inicio de nuevos proyectos.

Pros y contras de las acciones a tomar

Estudiaremos el potencial de despliegue de cada uno de los resultados obtenidos.

Estudiaremos si hay margen para la mejora de alguno de los procesos ejecutados hasta el momento.

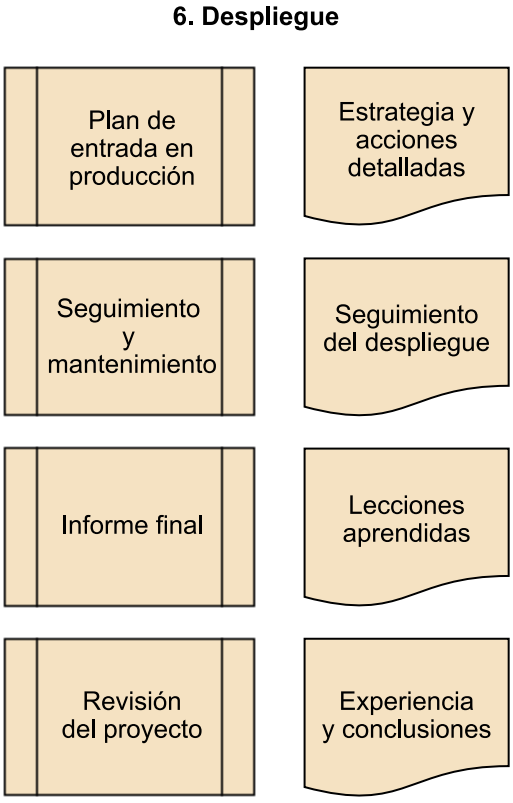
Argumentación de las decisiones tomadas

Realizaremos un ranking con las posibles acciones a tomar, seleccionaremos las más oportunas y argumentaremos nuestra decisión.

1.1.6. Despliegue

En esta fase se organizarán y ejecutarán tanto las tareas propias del despliegue de los resultados como del mantenimiento de las nuevas funcionalidades, una vez el despliegue haya finalizado.

Figura 8. Despliegue (CRISP-DM 1.0)



Plan de entrada en producción

El plan deberá contemplar todas las tareas a realizar en el proceso de despliegue de resultados, e incorporará medidas alternativas en forma de planes alternativos o versiones del plan inicial, que deberán permitir tener varias visiones y escoger la mejor.

Estrategia y acciones detalladas

Deberemos definir cómo el conocimiento obtenido en forma de resultados será propagado hacia los usuarios interesados.

En el caso de que haya que instalar o distribuir software por nuestros sistemas, deberemos gestionarlo para minimizar posibles efectos negativos y planificarlo para que se ejecute con suficiente antelación.

Habrá que prever cómo vamos a medir el beneficio producido por el despliegue y cómo vamos a monitorizar todo el proceso.

Identificaremos los posibles inconvenientes que pueda ocasionar nuestro despliegue.

Seguimiento y mantenimiento

Si el despliegue de los resultados del proyecto afecta a la actividad operativa de la organización, se hace imprescindible planificar y llevar a cabo tareas específicas de seguimiento y mantenimiento de las nuevas funcionalidades.

Seguimiento del despliegue

Deberemos tener en cuenta aspectos como el periodo de validez de los resultados o modelos desplegados, deberemos prever la monitorización de su grado de precisión.

Si los objetivos de negocio tienen fecha de caducidad o cambian con el paso del tiempo, parece lógico que los modelos propuestos tengan que adaptarse a esta situación.

En el caso de que el rendimiento del modelo esté por debajo de lo esperado, ¿qué deberemos hacer? Retirarlo del entorno productivo, retomar el proyecto para recalibrar el modelo, replantearse la solución propuesta e iniciar un proyecto más ambicioso,…

Informe final

Se puede plantear este informe final como un resumen de la experiencia del proyecto o como una presentación del proceso y de los resultados, lo adaptaremos en función del tipo de audiencia al que nos dirijamos.

Lecciones aprendidas

Plasmaremos en el informe final un resumen del trabajo realizado y de los logros conseguidos.

Adjuntaremos un estudio económico de los costes incurridos. Si es un proyecto interno, hablaremos de horas incurridas, de las desviaciones respecto de la previsión inicial.

Incluiremos también recomendaciones a tener en cuenta en futuros proyectos.

Revisión del proyecto

Evaluaremos las cosas que se han hecho bien y las que no se han hecho tan bien e identificaremos puntos y aspectos a mejorar.

Experiencia y conclusiones

Incluiremos entrevistas con los miembros del equipo de trabajo para conocer su visión.

Incluiremos si es posible entrevistas con usuarios finales en los que haya repercutido los resultados desplegados por el proyecto. ¿Mejorarían algo? ¿Reforzarían algún aspecto?

Haremos autocrítica constructiva identificando planteamientos equivocados que se hayan tenido que revisar o retomar.

Identificaremos las lecciones aprendidas a lo largo de las distintas fases del proyecto para incorporarlas como activos en futuros proyectos.

1.1.7. Objeciones a la metodología

La metodología que hemos presentado corresponde a la versión 1.0 de la misma y probablemente el grupo que la diseñó prepare mejoras y actualizaciones para versiones futuras.

Por ejemplo, faltaría dedicar tareas y actividades específicas al estudio del impacto del despliegue en la organización. Cada vez más las organizaciones responden a estructuras complejas de flujos de información, sistemas informáticos integrados e intereses miopes de departamentos o grupos de trabajo.

Un despliegue que no haya medido su impacto, a pesar de que sea claramente beneficioso y necesario, podría acarrear importantes desajustes en el equilibrio de lo que podríamos llamar el “ecosistema de la organización”.

En este sentido, se deberían organizar sesiones de trabajo y comunicación entre todos los departamentos implicados, formaciones a usuarios y prever recursos específicos para las tareas de despliegue, tanto a nivel de servicios centrales como a nivel de soporte *in situ* si fuera necesario.

1.2. Modelo DELTA para la mejora continua de BA

El modelo DELTA presentado por el autor Thomas H. Davenport en su libro *Competing on Analytics* no es por supuesto ni una metodología ni un estándar, pero hemos creído que merecía la pena presentarla en esta sección por su vocación de guía y por ser una colección muy acertada de pautas, tomadas desde una perspectiva de negocio y organizativa.

Gary Loveman, CEO de Harrah's Entertainment , Inc., reflexiona en el prefacio del libro.

> "Considero esencial cultivar en las organizaciones, de una forma sostenida en el tiempo, un ambiente en el que se separe claramente entre las ideas y las personas que las proponen, ya que este será el primer paso para poder distinguir con rigurosas evidencias entre las ideas que convienen ser llevadas a cabo y las que no".

Este ambiente de trabajo al que se refiere Gary Loveman, acompañado por un equipo de analistas tan bien formados como apasionados por su trabajo, es la clave para escalar en la pirámide de tipologías de organizaciones analíticas.

Ved también

Ved también el apartado "Estándar PMML"

Estudiaremos qué aspectos debe reforzar una organización para pasar al nivel superior en la pirámide de organización analítica. Para ello empezaremos por establecer las tres capacidades que una organización debe cuidar si quiere mejorar sus competencias analíticas:

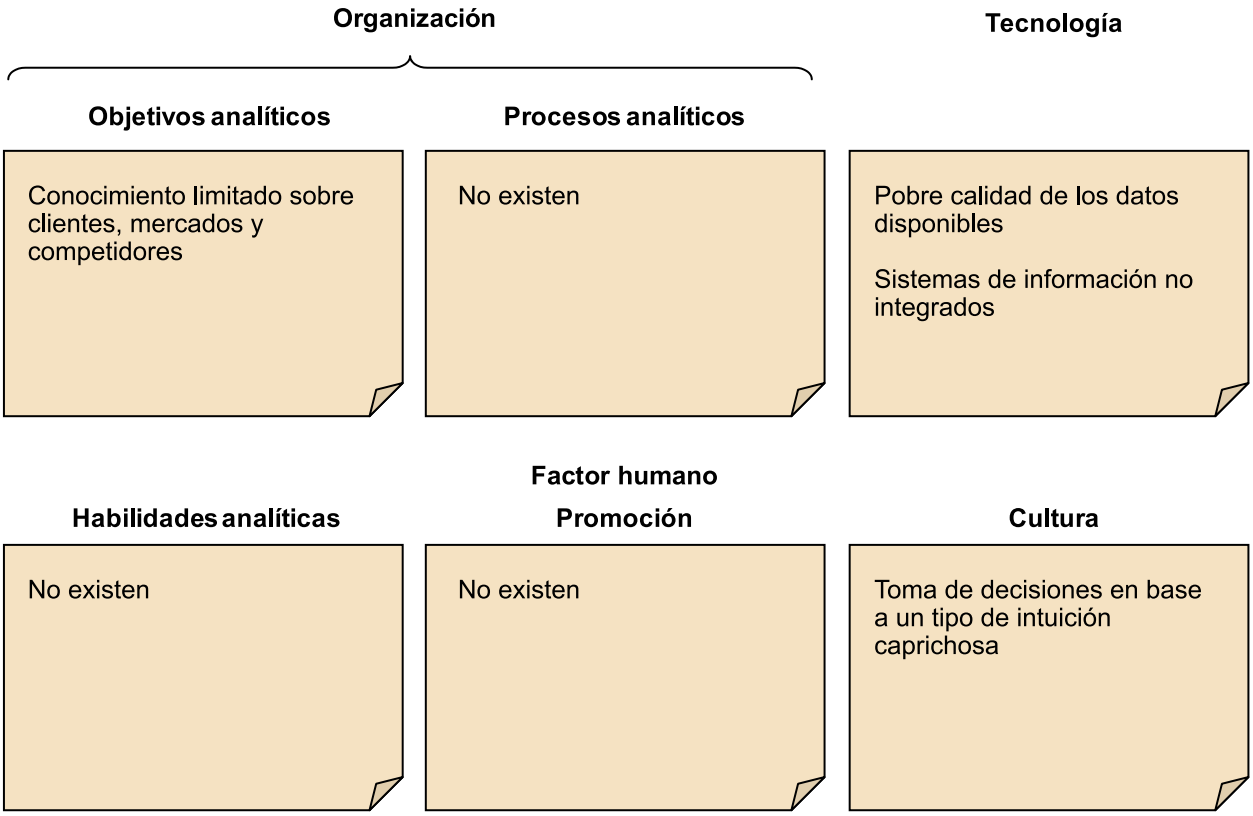
Organización – Recursos humanos – Tecnología

- Las capacidades organizativas, plasmadas en objetivos y procesos que soportan estos objetivos.
- Las capacidades humanas, que se concretan en habilidades, promoción de las actividades analíticas y cultura organizativa analítica.
- Las capacidades tecnológicas, que se refieren a disponer de datos de calidad y a disponer de aplicaciones integradas.

1.2.1. No consideran el análisis

Una organización que responde a la tipología de no considerar el análisis cumple con las siguientes características.

Figura 9. No consideran el análisis



Fuente: *Analytics at work, smarter decisions, better results*

El primer aspecto que se debería trabajar para poder tener cierta actividad analítica es el de disponer de un entorno transaccional que nos pueda garantizar una calidad en los datos referentes a las operaciones corrientes.

Asimismo, será muy conveniente poder disponer de alguna estrategia que nos permita focalizar la atención en aquellos datos que realmente son importantes para nuestra organización.

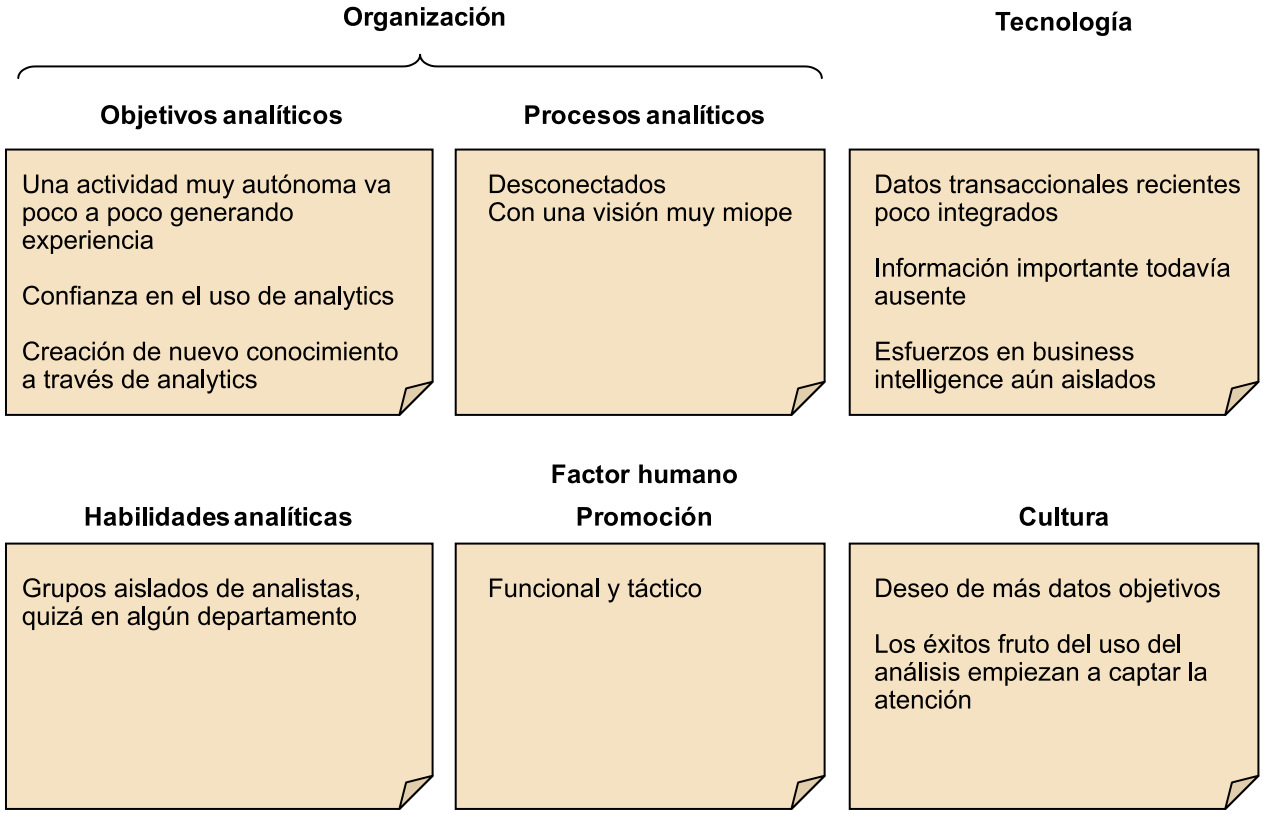
El siguiente paso sería realizar un estudio para evaluar de qué capacidades analíticas disponemos y en qué situación se encuentran. Habilidades de nuestros analistas, nivel de compromiso de la dirección, cultura de empresa respecto del análisis y el grado de promoción o aceptación de las actividades analíticas entre la fuerza de trabajo, son algunas de las capacidades de las que deberíamos tener una foto inicial.

Es normal que solo algunos departamentos o algunas actividades de la empresa estén en disposición de hacer el cambio analítico, incluso es posible que ya haya usuarios aislados o pequeños grupos de trabajo que estén llevando a cabo algún tipo de actividad analítica. A este respecto es importante identificarlos para posteriormente poder trabajar en su integración.

1.2.2. Actividad analítica aislada

Veamos primero cuál es el prototipo de organización que ya empieza a tener cierta actividad analítica y que por tanto cumple con lo que podríamos llamar los prerrequisitos de organización analítica.

Figura 10. Actividad analítica aislada



Fuente: *Analytics at work, smarter decisions, better results*

Para evolucionar, este tipo de organizaciones deben plantearse retos realistas, modestos, concretos y con un alcance muy acotado. Los pasos a seguir podrían ser los siguientes:

- Encontrar un patrocinador interno e identificar un *business problem*, es decir, un objetivo empresarial al que se le pueda dar respuesta mediante un pequeño proyecto de *business analytics*.

- Implementar un pequeño y localizado proyecto que contribuya a generar valor para la organización y que su beneficio sea cuantificable.

- Documentar los beneficios obtenidos y compartir y comunicar los resultados a los interesados.

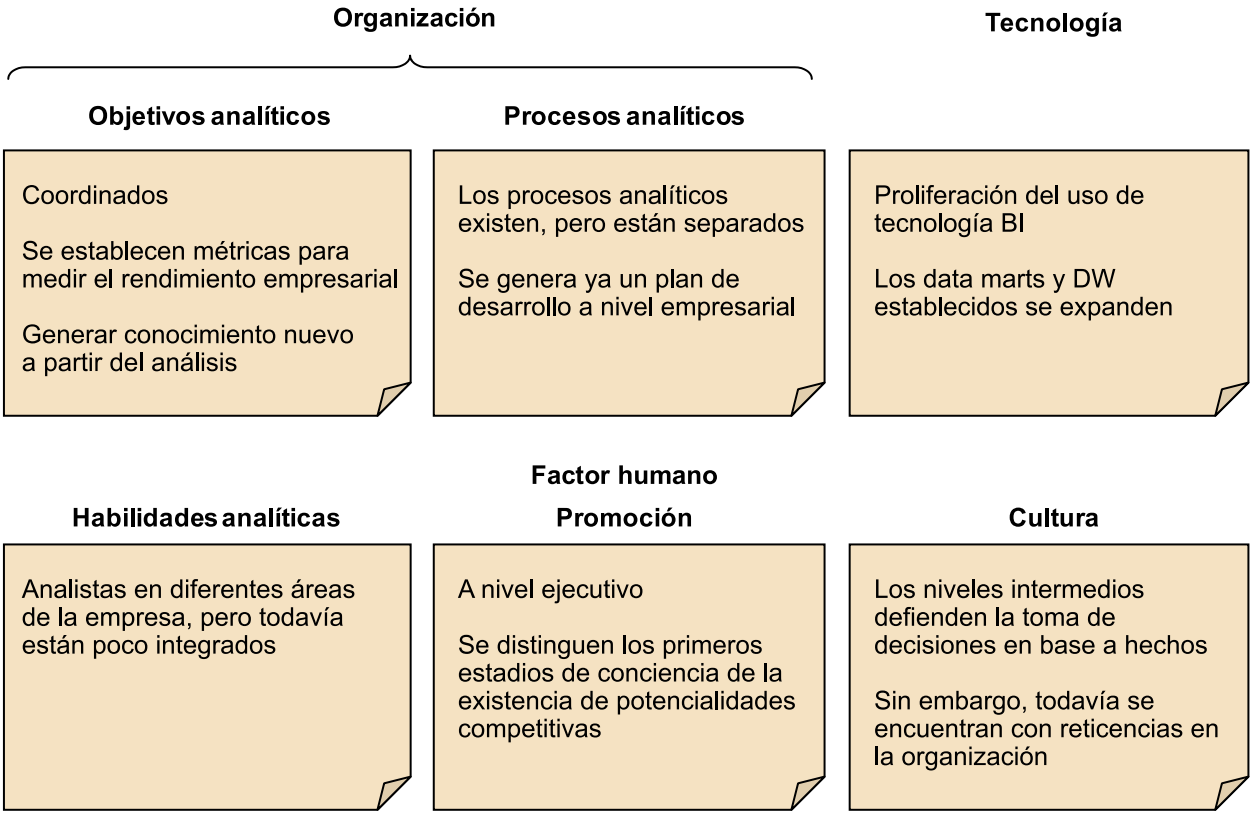
- Mantener una línea de construcción de pequeños proyectos que vayan generando beneficios para la organización, hasta que la propia organización haya ganado suficiente experiencia y se haya generado internamente su-

ficiente compromiso con los procesos analíticos como para poder dar, con garantías suficientes, el siguiente paso.

1.2.3. Aspirante analítico

La siguiente es la fotografía de una organización con perfil de aspirante analítico.

Figura 11. Aspirante analítico



Fuente: *Analytics at work, smarter decisions, better results*

Para llegar a este estadio, es vital contar con el apoyo de al menos una parte de la dirección de la organización, tanto es así que podríamos considerar que cualquier organización que cuente con este apoyo, al margen de cómo estén los otros factores, ya se puede considerar aspirante analítico.

El motivo es que con el compromiso de la dirección el resto de hitos serán alcanzables a corto y medio plazo. Esta debe proporcionar una visión estratégica de los objetivos que se pretenden alcanzar a partir de adoptar una cultura empresarial analítica.

Avanzar en esta línea significará disponer de una lista de métricas que nos deberá ayudar a medir el progreso de los procesos analíticos puestos en marcha.

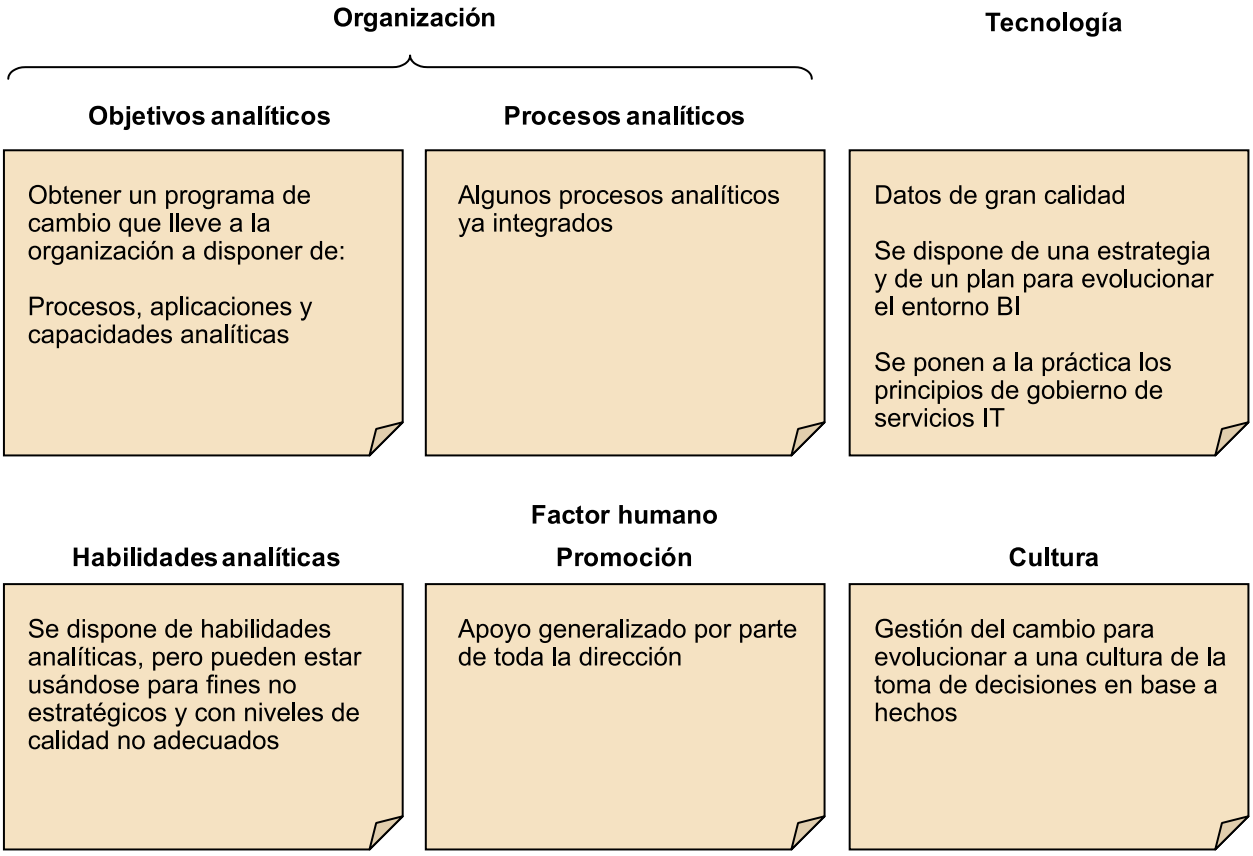
Será recomendable centralizar actividades analíticas para poder así profundizar más en la integración de procesos. Crear un centro de competencia de *business intelligence* puede ser un paso a dar.

La organización también deberá progresar en tecnología, incorporando aplicaciones que permitan a los analistas desarrollar sus capacidades, capacidades que habrá que gestionar y alinear en función de los objetivos marcados.

1.2.4. Organización analítica

La siguiente figura nos resume las capacidades de una organización analítica.

Figura 12. Organización analítica



Fuente: *Analytics at work, smarter decisions, better results*

En este estadio la promoción de la actividad analítica por parte de la dirección deja de estar en manos de algunos visionarios para pasar a ser un compromiso en firme por parte de toda la dirección en bloque y por extensión, para pasar a cualquier persona con responsabilidad de gestión.

Los retos en esta fase se multiplican y requieren de una profunda gestión del cambio para pasar a una cultura analítica y a una organización orientada a procesos analíticos integrados.

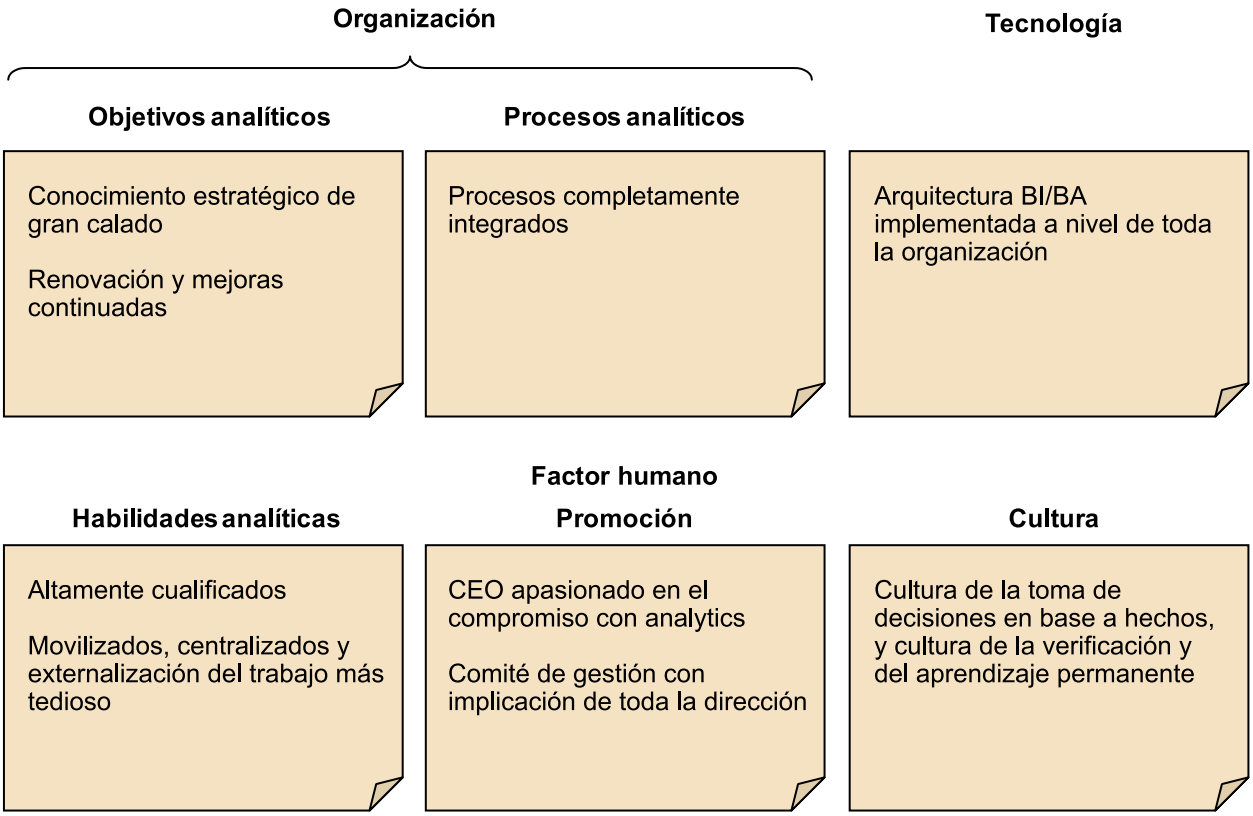
Este tipo de organizaciones suelen agrupar a los analistas más experimentados en centros de competencia claramente centrados en los aspectos más estratégicos para la organización.

La habilidad de desplegar unas capacidades analíticas sobresalientes sobre aquellos procesos que la organización considera críticos para su negocio, y hacerlo de forma integrada, es el reto que se plantea para obtener una organización analítica.

1.2.5. Competidor analítico

En lo alto de nuestra pirámide encontramos organizaciones que responden al siguiente patrón.

Figura 13. Competidor analítico



Fuente: Analytics at work, smarter decisions, better results

Para llegar a este estado se requiere que la habilidad analítica sea precisamente la principal ventaja competitiva de la empresa, la razón de ser de la misma, el producto estrella que justifica su presencia en el mercado.

El grado de excelencia debe ser de tal calibre que la convierta en innovadora en su campo, generando métricas, modelos analíticos o procesos analíticos propietarios y frutos de la investigación interna.

El reto consiste en mantener el nivel y para ello deberán evitar la autocomplacencia y establecer procesos que permanentemente miren al exterior en busca de signos y cambios en el mercado que obliguen a cambiar supuestos, modelos analíticos y reglas.

Podríamos enumerar una lista de empresas que se encuentran en este nivel de rendimiento, y seguramente cada una enfatiza más un aspecto que otro, o incluso habrán recorrido caminos distintos para llegar a este punto, pero todas ellas tienen en común…

…la creencia y apuesta por el mundo *business analytics*.

1.3. Estándar PMML

PMML es un lenguaje basado en XML y elaborado por the Data Mining Group, accesible en www.dmg.org, consorcio integrado por IBM, MicroStrategy, SAS y SPSS.

P r e d i c t i v e M o d e l M a r k u p L a n g u a g e

La razón de ser de PMML es la necesidad de disponer de un lenguaje estándar, es decir, compartido por la industria del software, para definir y compartir modelos de *data mining*.

De este modo, es posible, por ejemplo, utilizar una aplicación para generar el modelo y utilizar otra aplicación distinta para visualizarlo, analizarlo o evaluarlo, superando así las típicas barreras de incompatibilidades entre fabricantes.

La primera versión, la 0.7, fue liberada en julio de 1997 y la más reciente, la 4.1, es de diciembre del 2011. Se trata sin duda de una comunidad muy activa y de una apuesta clara de sus promotores para con la evolución del estándar.

La definición de los modelos se estructura en esquemas: cabecera, diccionarios de datos, transformaciones de datos, definición del modelo, esquema del modelo, objetivos y salidas. Se presenta un ejemplo práctico como anexo al material didáctico.

Cobertura del estándar PMML

A pesar de que la iniciativa está liderada por tres fabricantes, el estándar es aceptado por una lista mucho más extensa de empresas. A fecha del 2012 podemos encontrar las siguientes empresas adheridas:

Angoss, BlueLine, Business Objects, Crystal Ball, Dante, *data mining* Suite, DMG, DuckMiner, EMB, Experian, IBM, Info Centricity, Info Decipher, Insightful Miner, KNIME, KXEN, Laten View Analytics, Marketswitch Strategy

Tree Optimization, MatLab, Michael Decker, Microsoft, Micro Strategy, My Application, Open Tox, Oracle, Paragon, Pervasive, Rapid-I, Rattle, Rith Miner, Salford, SAP, SAS, Science Ops, SIPINA, SPSS, Star LIMS, Statsoft, Tree Net, Webtrends, Weka, Zementis.

Actualmente, PMML cubre 371 modelos, agrupados en un amplio abanico de familias de algoritmos *business analytics*:

Asociaciones, *clustering*, regresiones, *naive bayes*, redes neuronales, secuencias, *support vector machine*, series temporales, árboles de decisión y minería de textos.

1.4. Gobierno de servicios IT

No siempre, pero en muchas ocasiones, desplegar los resultados de un proyecto *business analytics* supondrá mejorar o crear nuevos servicios IT en nuestra organización. A pesar de haber trabajado ya con cierto detalle la metodología CRISP-DM que cubre las fases de despliegue y mantenimiento de los resultados de proyectos *data mining*, hemos considerado imprescindible saber qué dice la norma ISO20000 al respecto, como norma internacional de referencia para gobierno de servicios IT.

Hay que aclarar inicialmente que la ISO20000 no es ni un estándar ni una metodología, sino una norma o colección de buenas prácticas, cuyo objetivo es garantizar la prestación de servicios gestionados IT con una calidad aceptable para los clientes de un proveedor de servicios IT.

Estas buenas prácticas cumplen la condición de ser medibles, para posibilitar así la auditoría objetiva y posterior certificación del servicio IT auditado. Como siempre, recomendamos al estudiante que a la hora de ponerlas en práctica, tome en consideración la conveniencia de las mismas en función de las especificidades de su organización.

1.4.1. Definiciones

La ISO 20000 se crea en el 2005 como una evolución de la norma BS15000 y se redacta tomando como referencia las buenas prácticas de ITIL, muy centradas en la alineación de los servicios IT con las necesidades de la organización, y tomando también como referencia la norma ISO9000, muy orientada a la gestión de la calidad.

Hacer una lectura a las definiciones más relevantes que nos propone la ISO20000 nos ayudará mucho a comprender su filosofía.

Servicio

Es un sistema de información con soporte, que se entrega a un cliente con unos niveles de servicio previamente acordados.

Sistema de información

Es un sistema coherente de procesamiento de datos para el control o soporte de información en uno o más procesos de negocio. Está formado por personas, procesos y tecnología.

Proceso

Es un conjunto estructurado de actividades diseñado para cumplir un objetivo concreto.

Grupos de procesos

La ISO 20000 fomenta la adopción de un planteamiento de procesos integrados. Una organización que quiera desarrollar un sistema de gestión de la calidad tiene que identificar su propósito, definir las políticas y objetivos y determinar los procesos y su secuencia. Para planificar un proceso, una organización tiene que definir las actividades que componen el proceso.

Procedimiento

Un procedimiento es un documento que contiene los pasos que especifican cómo llevar a cabo una actividad. Los procedimientos están definidos como parte de procesos.

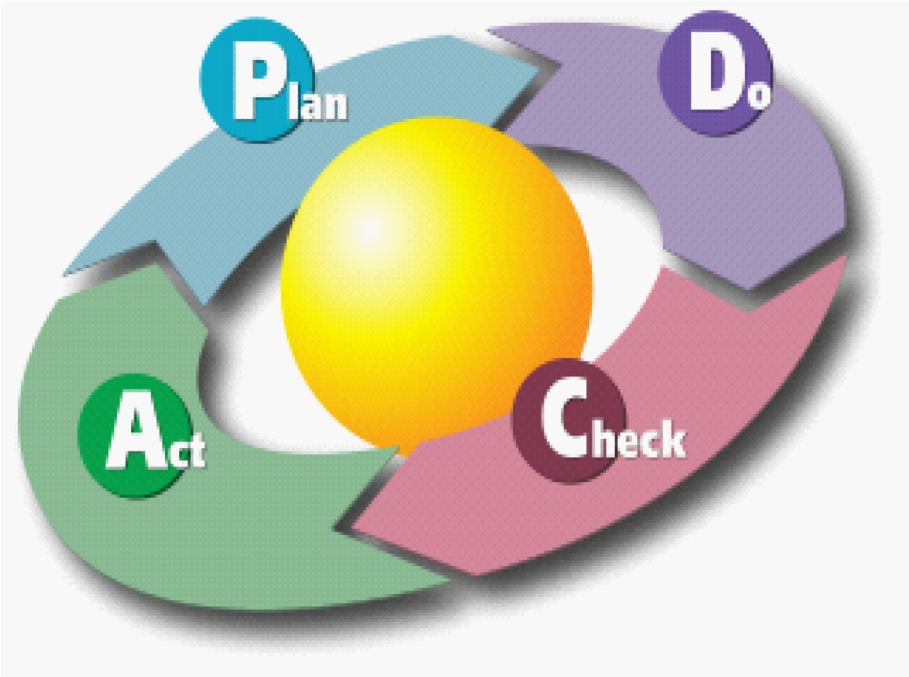
1.4.2. Procesos

Para la ISO20000, cada proceso debería diseñarse bajo un microciclo de gestión de la calidad PDCA.

Metodología PDCA, *Plan-Do-Check-Act*

La ISO20000 propone la metodología PDCA como estándar para gestionar cualquier proceso. Esta se basa en el círculo de calidad propuesto por Deming.

Figura 14. Círculo de Deming o círculo PDCA



- **Planificar**. Establecer los objetivos y los procesos necesarios para proporcionar resultados de acuerdo con las necesidades del cliente y con las políticas de la empresa.

- **Hacer**. Implementar los procesos.

- **Verificar**. Monitorizar y medir los procesos y los servicios contrastándolos con las políticas, los objetivos y los requisitos, e informar sobre los resultados.

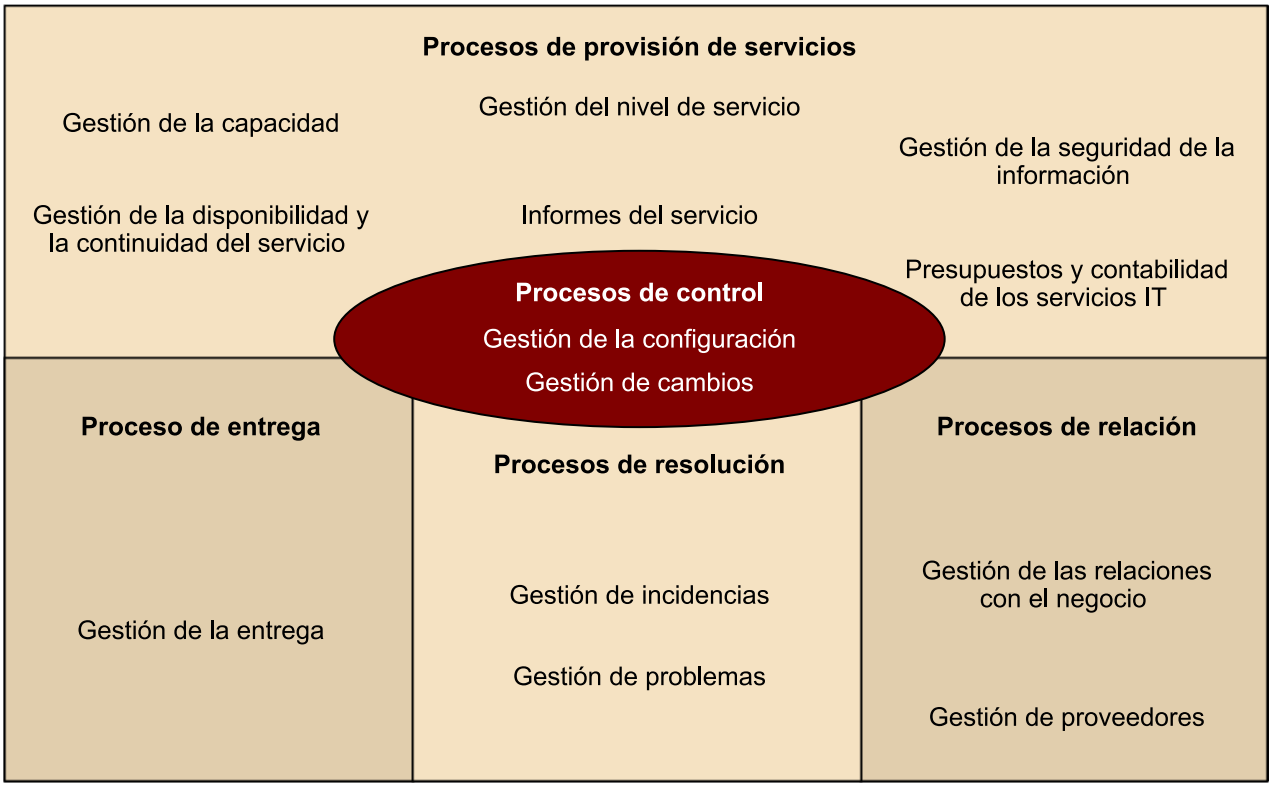
- **Actuar**. Emprender las acciones necesarias para mejorar continuamente el rendimiento y comportamiento del proceso.

Procesos de provisión de servicios

Detallamos a continuación los procesos que la ISO20000 distingue para la provisión de servicios, que recordemos, en el marco de *business analytics*, serían las mejoras de servicios existentes, como por ejemplo un cambio de criterio en el motor de recomendaciones de productos en nuestra tienda en línea, o bien la puesta en marcha de servicios nuevos, como podría ser la emisión de cupones descuento en los terminales de venta.

La ISO20000 nos permitirá tomar contacto con la importancia de planificar correctamente el impacto de nuestra actividad analítica en los sistemas ya en funcionamiento.

Figura 15. Gestión de servicios IT según la ISO20000



Los siguientes son los objetivos que la ISO20000 establece para cada uno de los procesos que identifica.

Gestión del nivel de servicio

Definir, acordar, registrar y gestionar los niveles de servicio.

Informes del servicio

Generar en plazo los informes acordados, fiables y precisos, que sirvan de apoyo a la toma de decisiones y faciliten una comunicación eficaz.

Gestión de la disponibilidad y la continuidad del servicio

Asegurar que los compromisos adquiridos con los clientes sobre la disponibilidad y la continuidad del servicio se pueden cumplir bajo todas las circunstancias.

Presupuestos y contabilidad de los servicios IT

Elaborar y controlar los presupuestos y contabilizar los costes de la provisión del servicio.

Gestión de la capacidad

Asegurar que el proveedor de servicio tiene, en todo momento, la capacidad suficiente para cubrir la demanda acordada, presente y futura, de las necesidades del negocio del cliente.

Gestión de la seguridad de la información

Gestionar eficazmente la seguridad de la información para todas las actividades del servicio.

Procesos de relación

Gestión de relaciones con el negocio

Establecer y mantener una buena relación entre el proveedor de servicios y el cliente, basándose en el entendimiento del cliente y de los fundamentos de su negocio.

Gestión de proveedores

Gestionar a los suministradores para garantizar la provisión sin interrupciones de servicios de calidad.

Procesos de resolución

Gestión de incidencias

Restaurar el servicio acordado con el negocio tan pronto como sea posible o responder a peticiones de servicio.

Gestión de problemas

Minimizar los efectos negativos sobre el negocio de interrupciones del servicio, mediante la identificación y el análisis proactivos de la causa de las incidencias y la gestión de los problemas para su cierre.

Procesos de control

Gestión de la configuración

Definir y controlar los componentes del servicio y de la infraestructura, y mantener información precisa sobre la configuración.

Gestión de los cambios

Asegurar que todos los cambios son evaluados, aprobados, implantados y revisados de una manera controlada.

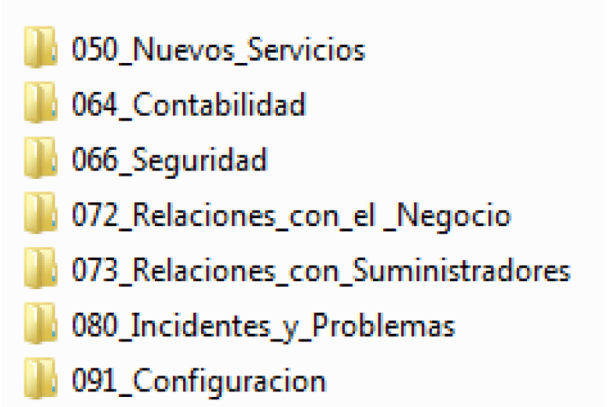
Procesos de entrega

La gestión de la entrega tiene por objetivo suministrar, distribuir y realizar el seguimiento de uno o más cambios en una entrega en el entorno de producción.

Resumiendo un poco, la ISO20000, combinada con una buena metodología de implementación de proyectos BA, debe ayudar a las organizaciones a poder crecer en funcionalidad analítica de una forma no traumática para ninguno de los actores involucrados, clientes, proveedores, usuarios, organizaciones y departamentos.

El grado de implicación de la organización con la norma puede ir desde simplemente organizar la documentación de gestión del centro de competencias analíticas, según aconseja la norma, como se ve en el ejemplo.

Figura 16. Resumen de procesos para la gestión de servicios IT



Hasta suponer una implementación rigurosa de la norma con el objetivo de superar una auditoría o conseguir una certificación.

2. *Data quality management*

Cuando realizamos tareas de *business analytics* con las que se busca obtener conocimiento a partir de los datos, parece más que obvio que los datos, como materia prima de nuestro proceso, deben ser de la máxima calidad.

¿Pero cómo podríamos definir el concepto de calidad en los datos?

Una primera aproximación sería observar que a una base de datos se le exige que cumpla con las propiedades de validez, la información falsa debe ser excluida, y de completitud, en el sentido de que no debe faltar información verdadera, es aquello de ...

> ..."toda la verdad (completitud) y nada más que la verdad (validez)"...
>
> Yan Zhang. *Noise Tolerant data mining*. The University of Vermont.

Esta definición ya nos da la idea de que la **veracidad** de los datos es la característica más importante en las bases de datos.

La **completitud** se refiere tanto a valores ausentes (campos vacíos) como a instancias ausentes. Por ejemplo, un estudio sobre el grado de aceptación de una leche para bebé, si contara tan solo con la opinión de los padres, sería incompleta, los profesionales sanitarios también deberían opinar.

Yendo más allá y teniendo en cuenta que en *business analytics* es frecuente que los datos provengan de diferentes fuentes, también será deseable evitar las **inconsistencias**.

Adicionalmente, hoy en día se da la circunstancia de que Internet, y las aplicaciones corporativas registran gran cantidad de datos transaccionales, con fecha de caducidad relativamente corta. De modo que se convierte en importante gestionar la **oportunidad** o validez temporal de los datos (*timeliness*).

De acuerdo con Kahn, Strong y Wang, 2002, datos de calidad se describen simplemente como...

> ..."datos que son apropiados para los consumidores de datos".

Estos mismos autores distinguen dos tipos de métricas para medir la calidad de los datos.

Las métricas objetivas y las métricas subjetivas, donde las primeras se basan exclusivamente en el propio juego de datos, mientras que las segundas requieren de la contextualización de los datos y del aporte de la experiencia del consumidor de datos.

Concretando un poco más podríamos clasificar las métricas en tres grupos:

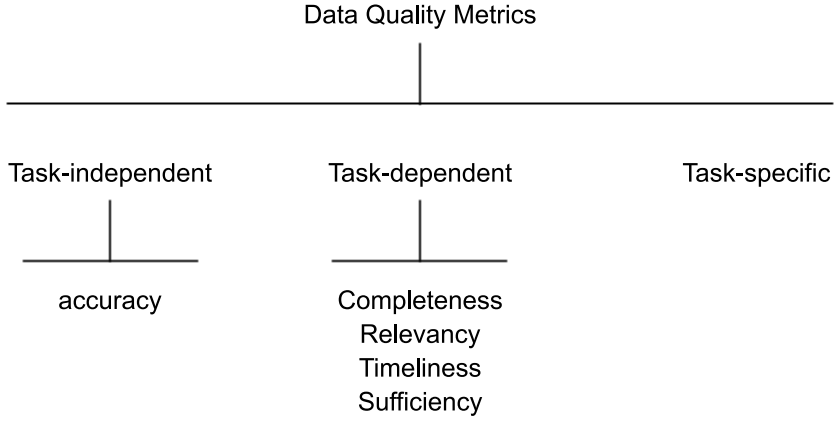
- Métricas **independientes de tarea**, serían las objetivas, donde lo importante es el dato en sí. Aquí encontramos el concepto de **precisión** (*accuracy*), que trata de medir la corrección del dato y asegurarse de que la fuente es fiable y el juego de datos imparcial.

- Métricas **dependientes de tarea**, serían las subjetivas, donde lo importante es colocar el dato en un contexto, en el uso que se le va a dar, en definitiva, la tarea que lo va a utilizar. Un juego de datos puede ser válido para un propósito y no serlo en absoluto para otro.
  En este ámbito se encuentra el concepto de **completitud**, que se ocupa de gestionar atributos, valores e instancias ausentes.
  La **relevancia** se ocupa de medir si un atributo es o no relevante para la tarea que lo precisa.
  La **oportunidad** o validez temporal, se ocupa de valorar si la información es suficientemente actual para el propósito de la tarea.
  La **suficiencia** mide si el volumen de datos disponible es apropiado para el objeto de la tarea.

- Métricas **específicas de tarea**, forman parte también de las subjetivas y se distinguen por servir solo para una tarea o ámbito concreto. Por ejemplo, en la industria farmacéutica un aspecto importante de la calidad de los datos incluye una representación válida de los ensayos clínicos.

La siguiente figura esquematiza la jerarquización de las métricas de calidad.

Figura 17. *Data quality metrics*



Pipino, Lee, y Wang, 2002

Si bien es cierto que la mayor parte de las tareas de gestión de la calidad de los datos suceden en la fase de preprocesado de la información, no es menos cierto que métricas específicas de tarea pueden ser necesarias en la misma fase de modelado.

Igualmente, merece la pena remarcar que en ocasiones, cuando tratamos de medir la precisión de los datos, debido a la dificultad que puede llegar a encerrar esta tarea, puede ser conveniente pasar de una visión objetiva del problema a una visión subjetiva del mismo, es decir, contextualizar los datos y hacer que sea el consumidor del dato quien valore su grado de precisión. Aclarar que el consumidor podría ser tanto una persona como un modelo o algoritmo.

Antes de entregar los datos a nuestro consumidor de datos, deberemos plantearnos la conveniencia de ejecutar algunas de las tareas que desarrollaremos a continuación. El objetivo de las mismas será garantizar que los datos entregados satisfacen las necesidades del consumidor de datos (calidad).

2.1. Preparación de los datos

En la fase de preparación de los datos ejecutaremos tareas como limpieza, integración, transformación y reducción.

Limpieza de datos (*data cleansing*)

En este ámbito se llevan a cabo actividades de detección, eliminación o corrección de instancias corrompidas o inapropiadas en los juegos de datos.

A nivel de valores de atributos se gestionan los valores ausentes, los erróneos (*outliers*), y los inconsistentes. Un ejemplo podrían ser los valores fuera de rango.

Integración de datos

Fruto de la fusión de juegos de datos distintos se pueden generar inconsistencias que deben ser detectadas y subsanadas.

Transformaciones

Se trata de actividades de manipulación de datos, como la normalización, discretización o la generalización de conceptos.

Reducción de datos

Se trata de técnicas que se aplican cuando el volumen de datos disponible es demasiado grande como para ser gestionado.

2.2. Discretización

La discretización es el proceso mediante el cual los valores de una variable continua se incluyen en contenedores, intervalos o grupos, con el objetivo de que haya un número limitado de estados posibles. Los contenedores se tratarán entonces como si fueran categorías.

La discretización es una tarea habitual en los procesos de minería de datos puesto que muchos algoritmos de clasificación requieren que sus atributos sean discretizados, bien porque solo acepten valores nominales, bien porque trabajar con atributos nominales en lugar de continuos disminuye el coste computacional y acelera el proceso inductivo.

Además, la discretización de atributos puede, en ocasiones, mejorar el rendimiento de un clasificador, como es el caso de C4.5.

Existen numerosos métodos para discretizar atributos. Veamos algunos.

Método de igual longitud

Disponemos de un atributo con n valores y queremos discretizarlo en k intervalos de igual longitud.

Sea $x_{\min}$ el valor mínimo que toma nuestro atributo y $x_{\max}$ el valor máximo. Entonces el intervalo será $\alpha = \frac{x_{\max} - x_{\min}}{k}$ y los límites $x_{\min} + i\alpha, \forall\, i = 1..k-1$

Este tipo de métodos no consideran el atributo objetivo o clase a la hora de establecer los intervalos. Este hecho provoca que en muchas ocasiones el propio proceso de discretización desencadene una pérdida de información.

Otros métodos más elaborados

El método 1RD propuesto por Holte (*machine learning*, 1994) trata de conseguir que cada intervalo solo contenga instancias de una clase o atributo objetivo.

Los métodos basados en la entropía pretenden dividir el dominio de valores en k intervalos de entropía mínima, es decir, se toma como criterio de discretización el valor que minimice la entropía de los subconjuntos generados utilizando ese punto de corte.

2.3. Gestión del ruido

Han y Kamber, 2000, nos definen ruido como…

> …"un error aleatorio o una variación en la medida de una variable".

Hickey 1996, para referirse al ruido, distingue entre atributos independientes y atributos objetivo (a predecir) en un juego de datos:

> "Ruido en los datos se refiere a cualquier cosa que oculta la relación entre los atributos independientes y los atributos objetivo".

Los factores que pueden ocasionar ruido en un juego de datos son varios:

- Un mal funcionamiento en equipos y aplicaciones.
- Errores en los procesos de transmisión de la información.
- Formularios rellenados por personas.
- La encriptación y desencriptación de datos puede generar errores.
- Asignación incorrecta de valores en una variable. En este caso distinguiremos si la variable es la variable objetivo o no.

Eliminación del ruido

La tarea de eliminar ruido a nivel de instancia, es decir, identificar las instancias erróneas, puede llevarse a cabo mediante algoritmos específicos para esta tarea o como parte de la propia fase de modelado.

Por ejemplo, en el algoritmo árbol de decisión C4.5, existe una tarea que es la poda del árbol. Esta puede llevarse a cabo en el propio proceso de construcción del árbol o al final del proceso de construcción como un postproceso.

La poda consiste simplemente en descartar aquellas instancias que se alejan demasiado del patrón habitual.

Otro ejemplo puede ser el uso del algoritmo clasificador K-Nearest Neighbor, específicamente para identificar errores, de modo que aquellas instancias que son incorrectamente clasificadas por el algoritmo pueden ser consideradas como errores (Tomek, 1976).

Siguiendo en el mismo ejemplo, Aha, Kibber y Albert, 1991, demostraron que seleccionar las instancias en función de su contribución a la precisión de la clasificación mejora mucho el rendimiento del modelo.

2.4. Reducción de la dimensionalidad

Respecto de la dimensionalidad de nuestro juego de datos, es importante estudiar el efecto de sus correlaciones, es decir, en un modelo predictivo nos interesará que las variables de entrada estén muy poco correlacionadas entre ellas y sin embargo, nos convendrá que las variables de entrada estén muy correlacionadas con la variable objetivo.

El grado de dependencia o independencia entre las variables de trabajo es un aspecto que hay que gestionar, puesto que muchos algoritmos contienen supuestos o prerrequisitos en este ámbito.

Al margen de las relaciones entre variables, también hay que tener en cuenta otras limitaciones que justifican la reducción de la dimensionalidad de los juegos de datos, una puede ser la propia capacidad de computación, que, a pesar que cada vez es mayor, no deja de ser un recurso con sus fronteras.

La otra limitación es la comprensibilidad del resultado, es decir, una salida de un modelo con excesivo número de variables puede dificultar a la mente humana la comprensión del resultado.

3. Anexo

3.1. Esquema PMML

Introducimos el esquema PMML mediante un ejemplo. El modelado de un juego de datos de varias cestas de la compra al que se le aplicará el algoritmo de asociaciones "Shopping Association SPSS".

A partir de 11 categorías de productos, se estudian 939 cestas de la compra, con el objetivo de encontrar asociaciones o implicaciones del estilo:

"si compra cerveza, también compra vegetales enlatados".

Presentamos a continuación una imagen resumida del código PMML que describe nuestro modelo, donde se distinguen los siguientes esquemas:

Cabecera

Es el esquema de presentación de la página PMML. Dispone de marcadores específicos para el *copyright* de la estructura PMML, un texto descriptivo del modelo, la aplicación software con la que se ha generado el modelo y la versión del software utilizado para generar el modelo.

Figura 18. PMML – Modelo de asociación– Cesta de la compra


```
▼<PMML xmlns="http://www.dmg.org/PMML-3_0"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="3.0">
 ▼<Header copyright="Copyright (c) Integral Solutions Ltd., 1994 - 2005.
  All rights reserved.">
    <Application name="Clementine" version="10.0"/>
   ▶<Annotation>...</Annotation>
  </Header>
 ▼<DataDictionary numberOfFields="2">
  ▼<DataField name="cardid" optype="continuous" dataType="integer">
     <Extension name="storageType" value="numeric"/>
   </DataField>
  ▼<DataField name="Product" optype="categorical" dataType="string">
     <Extension name="storageType" value="string"/>
     <Value value="beer" property="valid"/>
     <Value value="cannedmeat" property="valid"/>
     <Value value="cannedveg" property="valid"/>
     <Value value="confectionery" property="valid"/>
     <Value value="dairy" property="valid"/>
     <Value value="fish" property="valid"/>
     <Value value="freshmeat" property="valid"/>
     <Value value="frozenmeal" property="valid"/>
     <Value value="fruitveg" property="valid"/>
     <Value value="softdrink" property="valid"/>
     <Value value="wine" property="valid"/>
   </DataField>
  </DataDictionary>
 ▶<AssociationModel modelName="SHOPPING_ASSOC" algorithmName="Carma"
  functionName="associationRules" numberOfTransactions="939"
  minimumSupport="0.177848775292865" minimumConfidence="0.32013201320132"
  numberOfItems="7" numberOfItemsets="10"
  numberOfRules="18">...</AssociationModel>
</PMML>
```

Fuente: www.dmg.org

Diccionario de datos

Se define aquí el juego de datos que el modelo utilizará en el proceso de generación de asociaciones y reglas.

numberOfFields: Describe el número de atributos que componen nuestro diccionario. El ejemplo contiene 2 atributos: *cardid*, un numérico para identificar la compra y *product*, un categórico para identificar al producto comprado.

DataField: Contiene el identificador del atributo.

Optype: Es el tipo de campo en función de la operaciones que haremos sobre el mismo, el tipo categórico, nuestro ejemplo, solo puede ser comparado por "igual a". También podría ser de tipo ordinal, acepta una ordenación, o de tipo continuo, acepta operaciones aritméticas.

Value: Lista los valores aceptados por el campo que lo precede. Este marcador es sustitutivo del marcador *interval* que definiría rangos válidos.

Figura 19. PMML – Diccionario de datos – Modelo de asociación – Cesta de la compra

PMML – Diccionario de datos


```
▼<DataDictionary numberOfFields="2">
 ▼<DataField name="cardid" optype="continuous"
  dataType="integer">
    <Extension name="storageType" value="numeric"/>
  </DataField>
 ▼<DataField name="Product" optype="categorical"
  dataType="string">
    <Extension name="storageType" value="string"/>
    <Value value="beer" property="valid"/>
    <Value value="cannedmeat" property="valid"/>
    <Value value="cannedveg" property="valid"/>
    <Value value="confectionery" property="valid"/>
    <Value value="dairy" property="valid"/>
    <Value value="fish" property="valid"/>
    <Value value="freshmeat" property="valid"/>
    <Value value="frozenmeal" property="valid"/>
    <Value value="fruitveg" property="valid"/>
    <Value value="softdrink" property="valid"/>
    <Value value="wine" property="valid"/>
  </DataField>
</DataDictionary>
```

Fuente: www.dmg.org

dataType: Sirve para definir esquemas de información reutilizable, por ejemplo, podríamos definir un tipo de dato factura, que contendría un esquema formado por un número de factura, una fecha, una descripción,...

Esquema del modelo

Se define como la puerta de entrada al modelo, es decir, cualquier dato que pasemos al modelo debe cumplir con la estructura definida en el esquema.

MiningField: Declara y asigna propiedades a los campos que componen el esquema.

usageType: Define el papel que juega el campo dentro del esquema, los valores que puede tomar son:

- **active:** Atributo de entrada.
- **predicted:** Atributo objetivo del modelo. Atributo a predecir, en el caso de árboles de decisión.
- **supplementary:** Se trata de información adicional no necesaria para ejecutar el modelo. Podría usarse, por ejemplo, para explicar transformaciones anteriores.
- **outliers:** Define el método con el que gestionaremos los valores *outliers*. Las opciones disponibles son:

- **asIs:** Se mantiene el valor tal y como está. No se hace nada.

- **asMissingValues:** Se trata como si fuera un valor ausente o no informado.

- **asExtremeValues:** Los valores *outliers* se sobrescriben a un valor extremo por arriba o por abajo.

Figura 20. PMML – Esquema del modelo – Modelo de asociación – Cesta de la compra

PMML – Esquema del modelo

Fuente: www.dmg.org

Otras etiquetas importantes de este esquema definen cómo gestionar los valores ausentes o los valores no válidos.

Agrupaciones

Describe un método para representar el resultado del algoritmo asociativo. Se trata de un segmento específico de este tipo de modelos.

Itemset: Identifica una relación de valores del atributo sobre el que buscamos asociaciones.

numerOfItems: Número de valores que componen la agrupación o *itemset*.

Support: Soporte de la agrupación.

ItemRef: Identificador del valor del atributo "Producto". Hace referencia al id del "Item".

Figura 21. PMML – Agrupaciones soporte – Modelo de asociación – Cesta de la compra

PMML – Agrupaciones – Soporte

Fuente: www.dmg.org

Asociaciones

Lista todas las reglas de asociación identificadas en el juego de datos. Una regla consiste en una implicación y sus correspondientes medidas de soporte y esperanza.

Figura 22. PMML – Asociaciones reglas – Modelo de asociación – Cesta de la compra

PMML – Asociaciones - Reglas

Fuente: www.dmg.org

Presentamos a continuación otros esquemas no presentes en este ejemplo, pero que merece la pena mencionar.

Transformaciones

Es frecuente que los modelos precisen realizar transformaciones sobre los datos de entrada. Algunas de ellas pueden ser normalizaciones, discretizaciones, mapeo de valores, agregaciones o aplicación de funciones de conversión.

En este esquema se definen las transformaciones que se ejecutarán antes de aplicar el modelo.

Estadísticas

En ocasiones puede ser conveniente presentar ciertas estadísticas para variables que quizá no formen parte del modelado, pero que pueden ayudar a comprender o a soportar y verificar aspectos relevantes en el modelado.

Taxonomías y jerarquías

Los valores de una variable categórica puede ser que necesitemos representarlos de forma jerárquica, este es el caso por ejemplo de las familias de productos. Este esquema establece un entorno para llevar a cabo este tipo de representaciones.

Output

Se trata de elementos de salida que puede devolver el modelo. Estos elementos pueden ser simples variables con sus valores o incluso reglas con sus instrucciones para ser aplicadas.

Esta funcionalidad permite crear objetos PMML que actúen como consumidores de resultados de modelos, quizá para ejecutar nuevos modelos o para integrar los resultados en otros procesos.

Resumiendo, el estándar PMML es una herramienta que facilita enormemente los procesos de despliegue y mantenimiento de modelos en las organizaciones puesto que, literalmente, rompe la cadena que une el despliegue al desarrollo, separándolos en dos capas independientes y dotando así de flexibilidad a toda la parte más operativa de la minería de datos.

Resumen

En el capítulo dedicado a metodologías y estándares hemos desgranado todas las fases que nos propone CRISP-DM:

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación del modelo
- Despliegue

Aparentemente puede parecer, al igual que otras metodologías, algo pesada de seguir por exceso de detalle en sus procesos. Sin embargo, la ayuda de los esquemas resúmenes permite al estudiante retener una visión global de la metodología y le ayudan a situar las técnicas aprendidas en la fase y proceso más adecuados.

La posición de las organizaciones ante las posibilidades que les brinda un enfoque analítico, como el propuesto por BA, queda perfectamente visualizado mediante el modelo delta, que además nos da las pautas que una organización debería seguir para evolucionar su cultura organizativa hacia visiones más analíticas.

Hemos visto también cómo el estándar PMML surge como iniciativa de los principales fabricantes de software *data mining*, para dar respuesta a problemas de incompatibilidad entre productos y versiones distintas.

PMML permite, de una forma muy eficiente, que las organizaciones puedan cubrir sus necesidades de minería de datos, combinando software de distintos fabricantes.

BA, además de cubrir la clásica necesidad puntual de análisis y estudio de problemáticas concretas, también permite integrar "inteligencia" en procesos y servicios IT estables y permanentes. En este sentido, es importante saber qué dicen las buenas prácticas propuestas por la ISO 20000 respecto del gobierno de servicios IT.

De este modo, el estudiante se ha familiarizado con conceptos como la calidad total a través de la metodología PDCA, *Plan-Do-Check-Act*.

Por separado y en un capítulo específico, se ha trabajado la calidad de los datos, para, así, remarcar la importancia que en todo proyecto BA debe tener la completitud, consistencia y oportunidad de los datos.

Bibliografía

Davenport, T. H.; Harris, J.; Morison, R. (2010). "Analytics at Work: Smarter Decisions, Better Results". *Harvard Business Press*.

Davenport, T. H. (enero, 2006). "Competing on Analytics". *Harvard Business Review*.

Artículos

Davenport, T. H. "Analytics at Work: Q&A". www.informationweek.com/news/software/bi/222200096

Bisciglia, C. (2007). "Distributed Computing Seminar" (Lectura 4). Google.

Yan Zhang (mayo, 2008). *Noise tolerant Data Mining*. The University of Vermont.