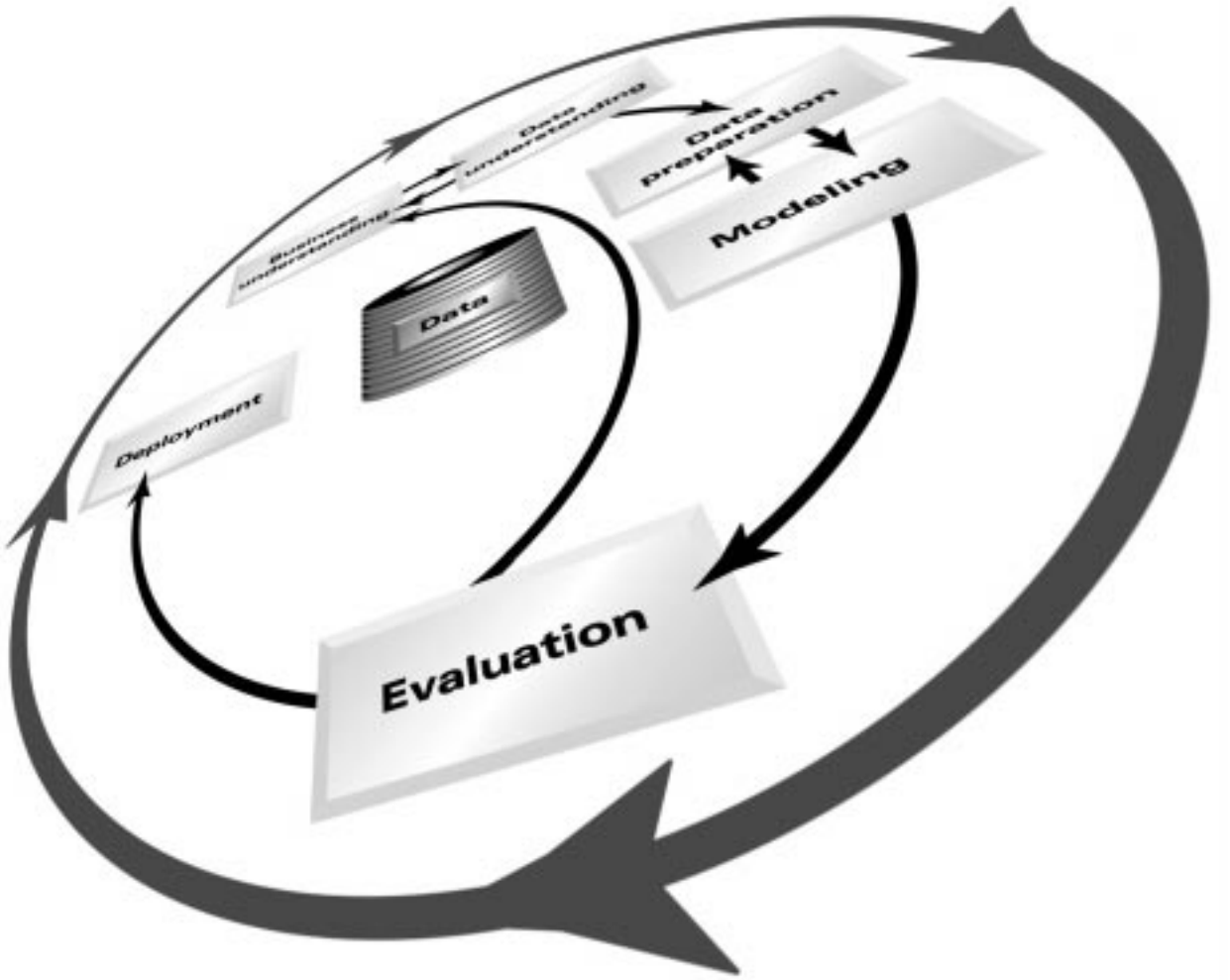
Traducido del inglés al español - www.onlinedoctranslator.com

CRISP-DM 1.0

Guía de minería de datos paso a paso



Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR),
Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler),
Colin Shearer (SPSS) y Rüdiger Wirth (DaimlerChrysler)

SPSS®

Este documento describe el modelo de proceso CRISP-DM, incluida una introducción a la metodología CRISP-DM, el modelo de referencia CRISP-DM, la guía del usuario CRISP-DM y los informes CRISP-DM, así como un apéndice con información adicional útil y relacionada. información.

Este documento y la información que contiene son propiedad exclusiva de los socios del consorcio CRISP-DM: NCR Systems Engineering Copenhagen (EE. UU. y Dinamarca), DaimlerChrysler AG (Alemania), SPSS Inc. (EE. UU.) y OHRA Verzekeringen en Bank Groep BV (Los países bajos)
Copyright © 1999, 2000

Todas las marcas comerciales y marcas de servicio mencionadas en este documento son marcas de sus respectivos propietarios y, como tales, son reconocidas por los miembros del consorcio CRISP-DM.

Prefacio

CRISP-DM fue concebido a fines de 1996 por tres "veteranos" del mercado de minería de datos joven e inmaduro. DaimlerChrysler (entonces Daimler-Benz) ya tenía experiencia, por delante de la mayoría de las organizaciones industriales y comerciales, en la aplicación de la minería de datos en sus operaciones comerciales. SPSS (entonces ISL) había brindado servicios basados en minería de datos desde 1990 y había lanzado el primer banco de trabajo de minería de datos comercial, Clementine, en 1994. NCR, como parte de su objetivo de brindar valor agregado a sus clientes de almacenamiento de datos de Teradata, había establecido equipos de consultores de minería de datos y especialistas en tecnología para atender los requerimientos de sus clientes.

En ese momento, el interés inicial del mercado en la minería de datos mostraba signos de explosión y una aceptación generalizada. Esto fue a la vez emocionante y aterrador. Todos nosotros habíamos desarrollado nuestros enfoques de minería de datos a medida que avanzábamos. ¿Lo estábamos haciendo bien? ¿Todos los nuevos adoptantes de la minería de datos tendrían que aprender, como lo hicimos inicialmente, por ensayo y error? Y desde la perspectiva de un proveedor, ¿cómo podríamos demostrarles a los clientes potenciales que la minería de datos estaba lo suficientemente madura como para ser adoptada como una parte clave de sus procesos comerciales?

Un modelo de proceso estándar, razonamos, no patentado y disponible gratuitamente, abordaría estos problemas para nosotros y para todos los profesionales.

Un año después formamos un consorcio, inventamos un acrónimo (CRoss-Industry Standard Process for Data Mining), obtuvimos financiación de la Comisión Europea y comenzamos a exponer nuestras ideas iniciales. Como CRISP-DM estaba destinado a ser neutral en cuanto a la industria, las herramientas y las aplicaciones, sabíamos que teníamos que obtener información de la mayor variedad posible de profesionales y otros (como proveedores de almacenamiento de datos y consultorías de gestión) con un interés personal en los datos. minería. Hicimos esto al crear el Grupo de Interés Especial CRISP-DM ("El SIG", como se le conoció). Lanzamos el SIG transmitiendo una invitación a las partes interesadas para que se unieran a nosotros en Ámsterdam para un taller de un día de duración: compartiríamos nuestras ideas, los invitaría a presentar las suyas y discutiríamos abiertamente cómo llevar adelante CRISP-DM.

El día del taller, había un sentimiento de inquietud entre los miembros del consorcio. ¿Nadie estaría lo suficientemente interesado como para presentarse? O si lo hicieran, ¿nos dirían que realmente no vieron una necesidad apremiante de un proceso estándar? ¿O que nuestras ideas estaban tan fuera de sintonía con las de los demás que cualquier idea de estandarización era una fantasía poco práctica?

El taller superó todas nuestras expectativas. Destacaban tres cosas:

- Asistió el doble de personas de lo que inicialmente esperábamos.
- Hubo un consenso abrumador de que la industria necesitaba un proceso estándar y lo necesitaba ahora.
- A medida que cada asistente presentó sus puntos de vista sobre la minería de datos a partir de su experiencia en proyectos, se hizo evidente que, aunque había diferencias superficiales, principalmente en la demarcación de las fases y en la terminología, había un gran terreno común en la forma en que veían el proceso de minería de datos.

Al final del taller, estábamos seguros de que podíamos entregar, con el aporte y la crítica del SIG, un modelo de proceso estándar para servir a la comunidad de minería de datos.

Durante los siguientes dos años y medio, trabajamos para desarrollar y perfeccionar CRISP-DM. Realizamos pruebas en proyectos de minería de datos a gran escala en vivo en Mercedes-Benz y en nuestro socio del sector de seguros, OHRA. Trabajamos en la integración de CRISP-DM con herramientas comerciales de minería de datos. El SIG demostró ser invaluable, creciendo a más de 200 miembros y realizando talleres en Londres, Nueva York y Bruselas.

Al final de la parte del proyecto financiada por la CE, a mediados de 1999, habíamos producido lo que consideramos un borrador de buena calidad del modelo de proceso. Quienes estén familiarizados con ese borrador encontrarán que un año después, aunque ahora mucho más completo y mejor presentado, CRISP-DM 1.0 no es radicalmente diferente. Éramos muy conscientes de que, durante el proyecto, el modelo de proceso todavía era un trabajo en progreso; CRISP-DM solo había sido validado en un conjunto limitado de proyectos. Durante el año pasado, DaimlerChrysler tuvo la oportunidad de aplicar CRISP-DM a una gama más amplia de aplicaciones. Los grupos de servicios profesionales de SPSS y NCR han adoptado CRISP-DM y lo han utilizado con éxito en numerosos compromisos con clientes que cubren muchas industrias y problemas comerciales. A lo largo de este tiempo, hemos visto a proveedores de servicios fuera del consorcio adoptar CRISP-DM; repetidas referencias a él por parte de los analistas como el estándar de facto para la industria; y una creciente conciencia de su importancia entre los clientes (ahora se hace referencia frecuente a CRISP-DM en las invitaciones a licitar y los documentos RFP). Creemos que nuestra iniciativa ha sido totalmente reivindicada y, si bien las futuras extensiones y mejoras son tanto deseables como inevitables, consideramos que la versión 1.0 de CRISP-DM está suficientemente validada para ser publicada y distribuida.

CRISP-DM no se ha construido de manera teórica y académica a partir de principios técnicos, ni los comités de élite de gurús lo crearon a puerta cerrada. Ambos enfoques para desarrollar metodologías se han probado en el pasado, pero rara vez han conducido a estándares prácticos, exitosos y ampliamente adoptados. CRISP-DM tiene éxito porque se basa sólidamente en la experiencia práctica del mundo real de cómo las personas realizan proyectos de minería de datos. Y en ese sentido, estamos abrumadoramente en deuda con los muchos profesionales que contribuyeron con sus esfuerzos e ideas a lo largo del proyecto.

El consorcio CRISP-DM

agosto de 2000

Contenido

Introducción

La metodología CRISP-DM

Desglose jerárquico

La metodología de minería de datos CRISP-DM se describe en términos de un modelo de proceso jerárquico, que consta de conjuntos de tareas descritas en cuatro niveles de abstracción (de general a específico): fase, tarea genérica, tarea especializada e instancia de proceso (ver figura 1). .

En el nivel superior, el proceso de minería de datos se organiza en varias fases; cada fase consta de varias tareas genéricas de segundo nivel. Este segundo nivel se llama genérico, porque pretende ser lo suficientemente general para cubrir todas las situaciones posibles de minería de datos. Las tareas genéricas pretenden ser lo más completas y estables posible. Completo significa que cubre tanto el proceso completo de minería de datos como todas las posibles aplicaciones de minería de datos. Estable significa que el modelo debe ser válido para desarrollos aún imprevistos, como nuevas técnicas de modelado.

El tercer nivel, el nivel de tareas especializadas, es el lugar para describir cómo se deben llevar a cabo las acciones de las tareas genéricas en ciertas situaciones específicas. Por ejemplo, en el segundo nivel puede haber una tarea genérica llamada limpieza de datos. El tercer nivel describe cómo esta tarea difería en diferentes situaciones, como la limpieza de valores numéricos frente a la limpieza de valores categóricos o si el tipo de problema es de agrupación o de modelado predictivo.

La descripción de fases y tareas como pasos discretos realizados en un orden específico representa una secuencia idealizada de eventos. En la práctica, muchas de las tareas se pueden realizar en un orden diferente y, a menudo, será necesario retroceder repetidamente a tareas anteriores y repetir ciertas acciones. Nuestro modelo de proceso no intenta capturar todas estas rutas posibles a través del proceso de minería de datos porque esto requeriría un modelo de proceso demasiado complejo.

El cuarto nivel, la instancia de proceso, es un registro de las acciones, decisiones y resultados de un compromiso real de minería de datos. Una instancia de proceso se organiza de acuerdo con las tareas definidas en los niveles superiores, pero representa lo que realmente sucedió en un compromiso particular, en lugar de lo que sucede en general.

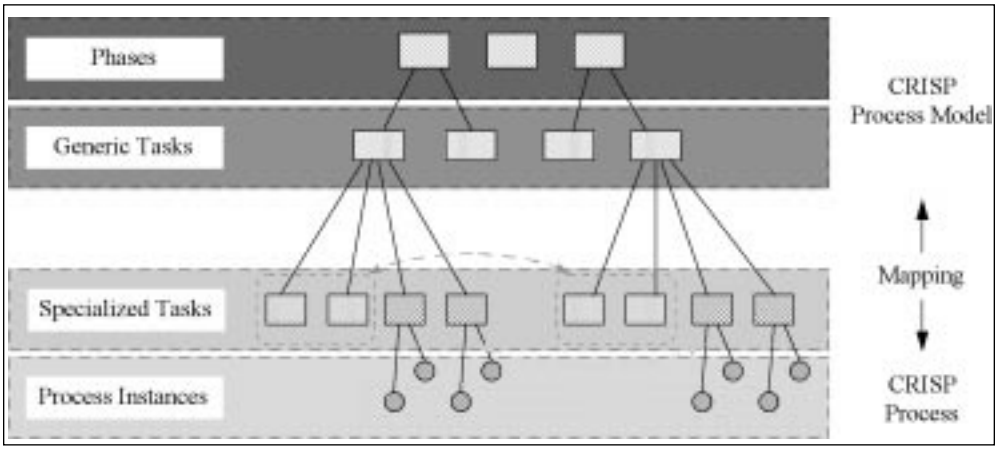


Figura 1: Desglose de cuatro niveles de la metodología CRISP-DM

Modelo de referencia y guía de usuario

Horizontalmente, la metodología CRISP-DM distingue entre el modelo de referencia y la guía del usuario. El modelo de referencia presenta una descripción general rápida de las fases, las tareas y sus resultados y describe qué hacer en un proyecto de minería de datos. La guía del usuario brinda consejos y sugerencias más detallados para cada fase y cada tarea dentro de una fase y describe cómo realizar un proyecto de minería de datos.

Este documento cubre tanto el modelo de referencia como la guía del usuario a nivel genérico.

Mapeo de modelos genéricos a modelos especializados

Contexto de minería de datos

El contexto de la minería de datos impulsa el mapeo entre el nivel genérico y el especializado en CRISP-DM. Actualmente, distinguimos entre cuatro dimensiones diferentes de contextos de minería de datos:

- El**dominio de aplicación**es el área específica en la que se lleva a cabo el proyecto de minería de datos.
- El**tipo de problema de minería de datos**describe la(s) clase(s) específica(s) de objetivo(s) que trata el proyecto de minería de datos (ver también el apéndice V.2).
- El**aspecto técnico**cubre problemas específicos en la minería de datos que describen diferentes desafíos (técnicos) que generalmente ocurren durante la minería de datos.
- El**herramienta y tecnica**La dimensión especifica qué herramientas y/o técnicas de minería de datos se aplican durante el proyecto de minería de datos.

La Tabla 1 a continuación resume estas dimensiones de los contextos de minería de datos y muestra ejemplos específicos para cada dimensión.

Tabla 1: Dimensiones de contextos y ejemplos de minería de datos

	Data Mining Context			
Dimension	*Application Domain*	*Data Mining Problem Type*	*Technical Aspect*	*Tool and Technique*
Examples	Response Modeling	Description and Summarization	Missing Values	Clementine
	Churn Prediction	Segmentation	Outliers	MineSet
	...	Concept Description	...	Decision Tree
		Classification		...
		Prediction		
		Dependency Analysis		

Un contexto de minería de datos específico es un valor concreto para una o más de estas dimensiones. Por ejemplo, un proyecto de minería de datos que se ocupa de un problema de clasificación en la predicción de abandono constituye un contexto específico. Cuantos más valores para diferentes dimensiones de contexto se fijan, más concreto es el contexto de minería de datos.

Mapeos con contextos

Distinguimos entre dos tipos diferentes de mapeo entre el nivel genérico y el especializado en CRISP-DM:

Mapeo para el presente:
Si solo aplicamos el modelo de proceso genérico para realizar un solo proyecto de minería de datos e intentamos asignar tareas genéricas y sus descripciones al proyecto específico según sea necesario, hablamos de un solo mapeo para (probablemente) solo un uso.

Mapeo para el futuro:
Si especializamos sistemáticamente el modelo de proceso genérico de acuerdo con un contexto predefinido (o, de manera similar, analizamos y consolidamos sistemáticamente las experiencias de un solo proyecto hacia un modelo de proceso especializado para uso futuro en contextos comparables), hablamos de escribir explícitamente un modelo de proceso especializado. en términos de CRISP-DM.

Qué tipo de mapeo es apropiado para sus propios propósitos depende de su contexto específico de minería de datos y las necesidades de su organización.

¿Cómo mapear?

La estrategia básica para mapear el modelo de proceso genérico al nivel especializado es la misma para ambos tipos de mapeos:

- Analiza tu contexto específico.
- Elimine cualquier detalle que no sea aplicable a su contexto.
- Agregue cualquier detalle específico a su contexto.
- Especialice (o instancia) contenidos genéricos según características concretas de su contexto.
- Posiblemente cambie el nombre de los contenidos genéricos para proporcionar significados más explícitos en su contexto en aras de la claridad.

Descripción de las piezas

Contenido

El modelo de proceso CRISP-DM (este documento) está organizado en cinco partes diferentes:

- La Parte I es esta introducción a la metodología CRISP-DM y proporciona algunas pautas generales para mapear el modelo de proceso genérico a modelos de proceso especializados.
- La Parte II describe el modelo de referencia CRISP-DM, sus fases, tareas genéricas y resultados.
- La Parte III presenta la guía del usuario de CRISP-DM que va más allá de la descripción pura de fases, tareas genéricas y resultados y contiene consejos más detallados sobre cómo realizar proyectos de minería de datos, incluidas listas de verificación.
- La Parte IV se centra en los informes que se producirán durante y después de un proyecto y sugiere esquemas para estos informes. También muestra referencias cruzadas entre productos y tareas.
- Finalmente, la parte V es el apéndice, que cubre un glosario de terminología importante, así como una caracterización de los tipos de problemas de minería de datos.

Objetivo

Los usuarios y lectores de este documento deben tener en cuenta las siguientes instrucciones:

- Si está leyendo el modelo de proceso CRISP-DM por primera vez, comience con la parte I, la introducción, para comprender la metodología CRISP-DM y todos sus conceptos y cómo los diferentes conceptos se relacionan entre sí. En lecturas posteriores, puede omitir la introducción y solo regresar a ella si es necesario para una aclaración.
- Si necesita acceso rápido a una descripción general del modelo de proceso CRISP-DM, consulte la parte II, el modelo de referencia CRISP-DM, ya sea para comenzar un proyecto de minería de datos rápidamente o para obtener una introducción a la guía del usuario de CRISP-DM.
- Si necesita asesoramiento detallado para realizar su proyecto de minería de datos, la parte III, la guía del usuario de CRISP-DM, es la parte más valiosa de este documento. Nota: si no ha leído primero la introducción o el modelo de referencia, retroceda y comience a leer con estas dos primeras partes.

- Si se encuentra en la etapa de extracción de datos cuando escribe sus informes, salte a la parte IV. Si prefiere generar sus descripciones de entregables durante el proyecto, avance y retroceda entre las partes III y IV según lo desee.
- Finalmente, el apéndice es útil como información básica adicional sobre CRISP-DM y la minería de datos. Use el apéndice para buscar varios términos si aún no es un experto en el campo.

II El modelo de referencia CRISP-DM

El modelo de proceso actual para la minería de datos proporciona una descripción general del ciclo de vida de un proyecto de minería de datos. Contiene las fases de un proyecto, sus respectivas tareas y las relaciones entre estas tareas. En este nivel de descripción, no es posible identificar todas las relaciones. Esencialmente, podrían existir relaciones entre cualquier tarea de minería de datos según los objetivos, los antecedentes y el interés del usuario y, lo que es más importante, de los datos.

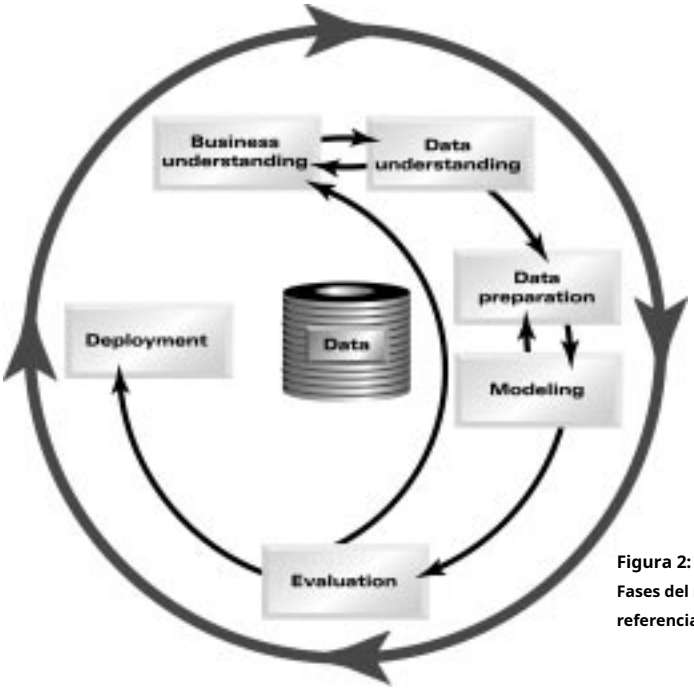


Figura 2:
Fases del modelo de referencia CRISP-DM

El ciclo de vida de un proyecto de minería de datos consta de seis fases. La figura 2 muestra las fases de un proceso de minería de datos. La secuencia de las fases no es rígida. Siempre es necesario avanzar y retroceder entre las diferentes fases. Depende del resultado de cada fase qué fase o qué tarea particular de una fase debe realizarse a continuación. Las flechas indican las dependencias más importantes y frecuentes entre fases.

El círculo exterior en la Figura 2 simboliza la naturaleza cíclica de la minería de datos en sí. La minería de datos no termina una vez que se implementa una solución. Las lecciones aprendidas durante el proceso y de la solución implementada pueden desencadenar nuevas preguntas comerciales, a menudo más enfocadas. Los procesos posteriores de minería de datos se beneficiarán de las experiencias de los anteriores.

A continuación, describimos brevemente cada fase:

Comprensión empresarial
Esta fase inicial se enfoca en comprender los objetivos y requisitos del proyecto desde una perspectiva comercial, y luego convertir este conocimiento en una definición del problema de minería de datos y un plan preliminar diseñado para lograr los objetivos.

Comprensión de datos

La fase de comprensión de datos comienza con una recopilación inicial de datos y continúa con actividades para familiarizarse con los datos, identificar problemas de calidad de datos, descubrir los primeros conocimientos sobre los datos o detectar subconjuntos interesantes para formar hipótesis sobre información oculta.

Preparación de datos

La fase de preparación de datos cubre todas las actividades para construir el conjunto de datos final (datos que se incorporarán a la(s) herramienta(s) de modelado) a partir de los datos sin procesar iniciales. Es probable que las tareas de preparación de datos se realicen varias veces y no en ningún orden prescrito. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de datos para herramientas de modelado.

Modelado

En esta fase, se seleccionan y aplican varias técnicas de modelado y sus parámetros se calibran a valores óptimos. Por lo general, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos específicos sobre la forma de los datos. Por lo tanto, a menudo es necesario retroceder a la fase de preparación de datos.

Evaluación

En esta etapa del proyecto, ha creado un modelo (o modelos) que parecen tener una alta calidad desde la perspectiva del análisis de datos. Antes de proceder a la implementación final del modelo, es importante evaluar más a fondo el modelo y revisar los pasos ejecutados para construir el modelo para asegurarse de que logre correctamente los objetivos comerciales. Un objetivo clave es determinar si existe algún tema comercial importante que no se haya considerado lo suficiente. Al final de esta fase, se debe llegar a una decisión sobre el uso de los resultados de la minería de datos.

Despliegue

La creación del modelo generalmente no es el final del proyecto. Incluso si el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento obtenido deberá organizarse y presentarse de manera que el cliente pueda utilizarlo. Suele implicar la aplicación de modelos "vivos" dentro de los procesos de toma de decisiones de una organización, por ejemplo, en la personalización en tiempo real de páginas web o la puntuación repetida de bases de datos de marketing. Sin embargo, según los requisitos, la fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso de extracción de datos repetible en toda la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva a cabo los pasos de implementación. Sin embargo,

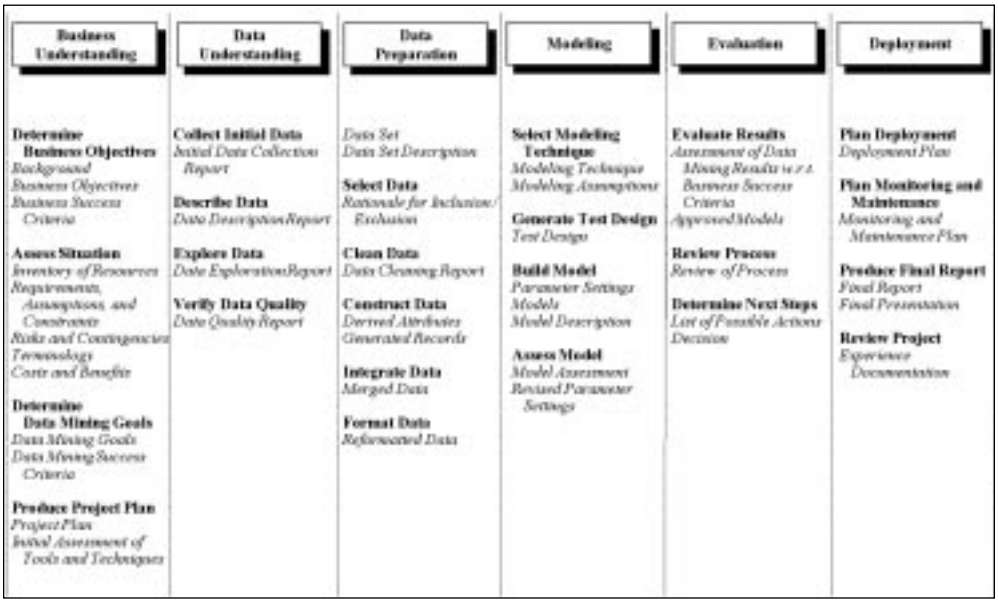


Figura 3: Tareas genéricas (negrita) y resultados (cursiva) del modelo de referencia CRISP-DM

La Figura 3 presenta un esquema de fases acompañado de tareas genéricas (negrita) y productos (cursiva). En las siguientes secciones, describimos cada tarea genérica y sus resultados con más detalle. Enfocamos nuestra atención en resúmenes de tareas y resúmenes de resultados.

1 Comprensión empresarial

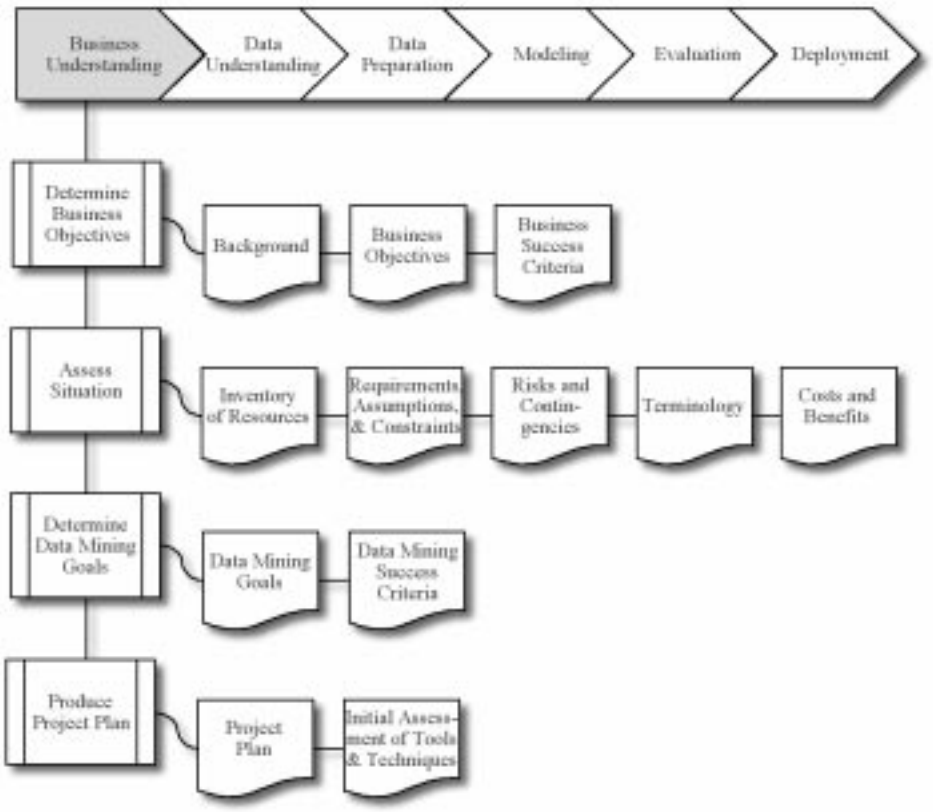


Figura 4: Comprensión empresarial

1.1 Determinar los objetivos comerciales

Tarea

Determinar los objetivos comerciales.

El primer objetivo del analista de datos es comprender a fondo, desde una perspectiva comercial, lo que el cliente realmente quiere lograr. A menudo, el cliente tiene muchos objetivos y limitaciones en competencia que deben equilibrarse adecuadamente. El objetivo del analista es descubrir factores importantes, al principio, que pueden influir en el resultado del proyecto. Una posible consecuencia de descuidar este paso es gastar mucho esfuerzo en producir las respuestas correctas a las preguntas incorrectas.

Salidas

Fondo

Registrar la información que se conoce sobre la situación empresarial de la organización al inicio del proyecto.

Objetivos de negocios

Describa el objetivo principal del cliente, desde una perspectiva empresarial. Además del objetivo comercial principal, normalmente hay otras preguntas comerciales relacionadas que al cliente le gustaría abordar. Por ejemplo, el objetivo comercial principal podría ser mantener a los clientes actuales mediante la predicción de cuándo son propensos a pasarse a un competidor. Ejemplos de preguntas comerciales relacionadas son "¿Cómo afecta el canal principal (p. ej., cajero automático, visita a una sucursal, Internet) que usa un cliente del banco si se queda o se va?" o "¿La reducción de las tarifas de los cajeros automáticos reducirá significativamente la cantidad de clientes de alto valor que se van?"

Criterios de éxito empresarial

Describa los criterios para un resultado exitoso o útil para el proyecto desde el punto de vista empresarial. Esto puede ser bastante específico y puede medirse objetivamente, como la reducción de la rotación de clientes a un cierto nivel o general y subjetivo, como "brindar información útil sobre las relaciones". En este último caso se debe indicar quién hace el juicio subjetivo.

1.2 Evaluar la situación

Tarea **evaluar la situación**

Esta tarea implica una búsqueda de hechos más detallada sobre todos los recursos, restricciones, suposiciones y otros factores que se deben considerar para determinar el objetivo del análisis de datos y el plan del proyecto. En la tarea anterior, tu objetivo es llegar rápidamente al quid de la situación. Aquí, desea desarrollar los detalles.

Salidas

Inventario de recursos

Enumere los recursos disponibles para el proyecto, incluidos: personal (expertos comerciales, expertos en datos, soporte técnico, personal de extracción de datos), datos (extractos fijos, acceso a datos operativos o almacenados en vivo), recursos informáticos (plataformas de hardware) y software (recursos de datos). herramientas de minería, otro software relevante).

Requisitos, suposiciones y restricciones

Enumere todos los requisitos del proyecto, incluido el calendario de finalización, la comprensibilidad y la calidad de los resultados y la seguridad, así como las cuestiones legales. Como parte de este resultado, asegúrese de que tiene permiso para usar los datos.

Enumere las suposiciones hechas por el proyecto. Estas pueden ser suposiciones sobre los datos que se pueden verificar durante la extracción de datos, pero también pueden incluir suposiciones no verificables sobre el negocio en el que se basa el proyecto. Es particularmente importante enumerar estos últimos si forman condiciones sobre la validez de los resultados.

Enumere las restricciones del proyecto. Estas pueden ser restricciones en la disponibilidad de recursos, pero también pueden incluir restricciones tecnológicas como el tamaño de los datos que es práctico usar para el modelado.

Riesgos y contingencias

Enumere los riesgos o eventos que podrían ocurrir para retrasar el proyecto o hacer que fracase. Listar los planes de contingencia correspondientes; qué medidas se tomarán si se presentan los riesgos.

Terminología

Compilar un glosario de terminología relevante para el proyecto. Esto puede incluir dos componentes:

(1) Un glosario de terminología comercial relevante, que forma parte de la comprensión comercial disponible para el proyecto. La construcción de este glosario es un ejercicio útil de "obtención de conocimientos" y educación.

(2) Un glosario de terminología de minería de datos, ilustrado con ejemplos relevantes para el problema comercial en cuestión.

Costos y beneficios

Construya un análisis de costo-beneficio para el proyecto, que compare los costos del proyecto con el beneficio potencial para el negocio si tiene éxito. La comparación debe ser lo más específica posible, por ejemplo, utilizando medidas monetarias en una situación comercial.

1.3 Determinar los objetivos de la minería de datos

Tarea

Determinar los objetivos de minería de datos

A*objetivo de negocio*establece los objetivos en la terminología empresarial. A*objetivo de minería de datos*establece los objetivos del proyecto en términos técnicos. Por ejemplo, el objetivo comercial podría ser "Aumentar las ventas por catálogo a los clientes existentes". Un objetivo de minería de datos podría ser "Predecir cuántos widgets comprará un cliente, dadas sus compras en los últimos tres años, la información demográfica (edad, salario, ciudad, etc.) y el precio del artículo".

Salidas

Objetivos de minería de datos

Describa los resultados previstos del proyecto que permiten el logro de los objetivos comerciales.

Criterios de éxito de la minería de datos

Defina los criterios para un resultado exitoso del proyecto en términos técnicos, por ejemplo, un cierto nivel de precisión predictiva o un perfil de propensión a comprar con un grado determinado de "elevación". Al igual que con los criterios de éxito comercial, puede ser necesario describirlos en términos subjetivos, en cuyo caso se debe identificar a la persona o personas que realizan el juicio subjetivo.

1.4 Producir plan de proyecto

Tarea

Producir el plan de proyecto

Describa el plan previsto para lograr los objetivos de minería de datos y, por lo tanto, lograr los objetivos comerciales. El plan debe especificar el conjunto anticipado de pasos a realizar durante el resto del proyecto, incluida una selección inicial de herramientas y técnicas.

Salidas

Plan de proyecto

Enumere las etapas a ejecutar en el proyecto, junto con la duración, los recursos requeridos, las entradas, las salidas y las dependencias. Siempre que sea posible, haga explícitas las iteraciones a gran escala en el proceso de extracción de datos, por ejemplo, repeticiones de las fases de modelado y evaluación.

Como parte del plan del proyecto, también es importante analizar las dependencias entre el cronograma y los riesgos. Marque los resultados de estos análisis explícitamente en el plan del proyecto, idealmente con acciones y recomendaciones si aparecen los riesgos.

Nota: el plan del proyecto contiene planes detallados para cada fase. Por ejemplo, decida en este punto qué estrategia de evaluación se utilizará en la fase de evaluación.

El plan del proyecto es un documento dinámico en el sentido de que al final de cada fase es necesaria una revisión del progreso y los logros y, en consecuencia, se recomienda una actualización del plan del proyecto. Los puntos de revisión específicos para estas revisiones también forman parte del plan del proyecto.

Evaluación inicial de herramientas y técnicas.

Al final de la primera fase, el proyecto también realiza una evaluación inicial de herramientas y técnicas. Aquí, selecciona una herramienta de minería de datos que admita varios métodos para diferentes etapas del proceso, por ejemplo. Es importante evaluar las herramientas y técnicas al principio del proceso, ya que la selección de herramientas y técnicas posiblemente influya en todo el proyecto.

2 Comprensión de datos

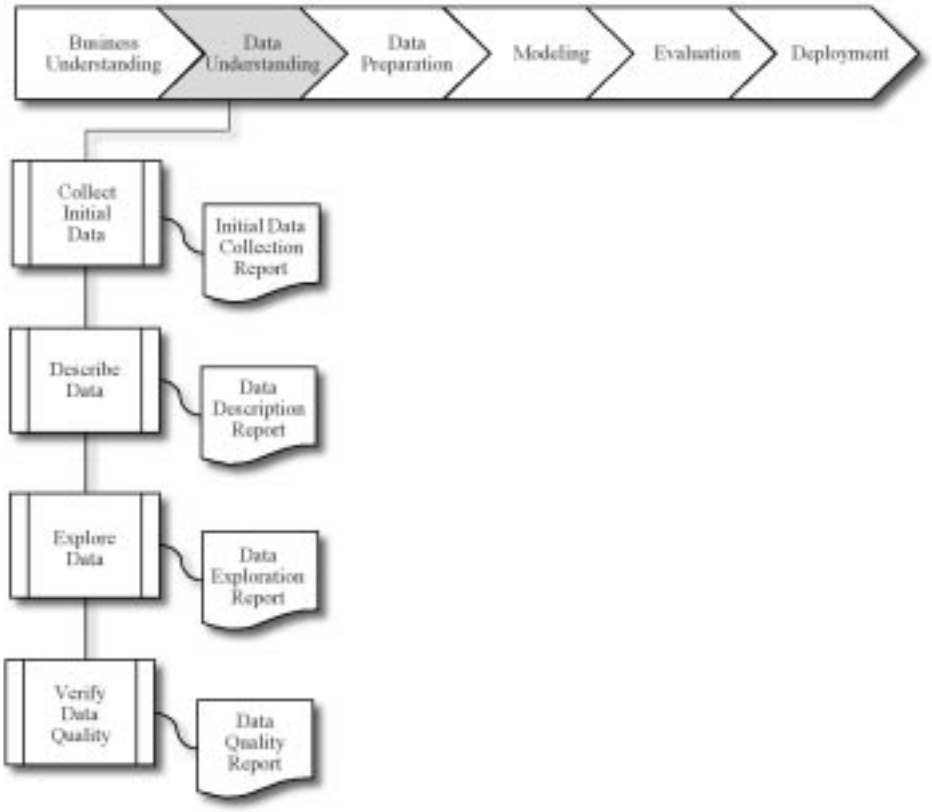


Figura 5: Comprensión de datos

2.1 Recolectar datos iniciales

Tarea **Recopilar datos iniciales**

Adquirir dentro del proyecto los datos (o acceder a los datos) enumerados en los recursos del proyecto. Esta recopilación inicial incluye la carga de datos si es necesario para la comprensión de los datos. Por ejemplo, si aplica una herramienta específica para la comprensión de datos, tiene mucho sentido cargar sus datos en esta herramienta. Este esfuerzo posiblemente conduzca a los pasos iniciales de preparación de datos.

Nota: si adquiere varias fuentes de datos, la integración es un problema adicional, ya sea aquí o en la fase posterior de preparación de datos.

Producción **Informe inicial de recopilación de datos**

Enumere el conjunto de datos (o conjuntos de datos) adquiridos, junto con sus ubicaciones dentro del proyecto, los métodos utilizados para adquirirlos y cualquier problema encontrado. Registre los problemas encontrados y cualquier solución lograda para ayudar con la replicación futura de este proyecto o con la ejecución de proyectos futuros similares.

2.2 Describir datos

Tarea **Describir datos**

Examine las propiedades "brutas" o "superficiales" de los datos adquiridos e informe sobre los resultados.

Producción **Informe de descripción de datos**

Describa los datos que se han adquirido, incluidos: el formato de los datos, la cantidad de datos, por ejemplo, el número de registros y campos en cada tabla, las identidades de los campos y cualquier otra característica superficial de los datos que se han descubierto. ¿Los datos adquiridos cumplen los requisitos pertinentes?

2.3 Explorar datos

Tarea **Explorar datos**

Esta tarea aborda las preguntas de minería de datos, que se pueden abordar mediante consultas, visualización e informes. Estos incluyen: distribución de atributos clave, por ejemplo, el atributo objetivo de una tarea de predicción; relaciones entre pares o pequeños números de atributos; resultados de agregaciones simples; propiedades de subpoblaciones significativas; análisis estadísticos simples. Estos análisis pueden abordar directamente los objetivos de minería de datos; también pueden contribuir o refinar la descripción de datos y los informes de calidad y alimentar la transformación y otra preparación de datos necesaria para un análisis posterior.

Producción **Informe de exploración de datos**

Describa los resultados de esta tarea, incluidos los primeros hallazgos o la hipótesis inicial y su impacto en el resto del proyecto. Si corresponde, incluya gráficos y diagramas que indiquen las características de los datos o conduzcan a subconjuntos de datos interesantes para un examen más detallado.

2.4 Verificar la calidad de los datos

Tarea

Verificar la calidad de los datos

Examinar la calidad de los datos, abordando preguntas como: ¿están completos los datos (cubren todos los casos requeridos)? ¿Es correcto o contiene errores y, si hay errores, qué tan comunes son? ¿Hay valores faltantes en los datos? Si es así, ¿cómo se representan, dónde ocurren y qué tan comunes son?

Producción

Informe de calidad de datos

Listar los resultados de la verificación de la calidad de los datos; si existen problemas de calidad, enumere las posibles soluciones. Las soluciones a los problemas de calidad de los datos generalmente dependen en gran medida tanto de los datos como del conocimiento comercial.

3 Preparación de datos

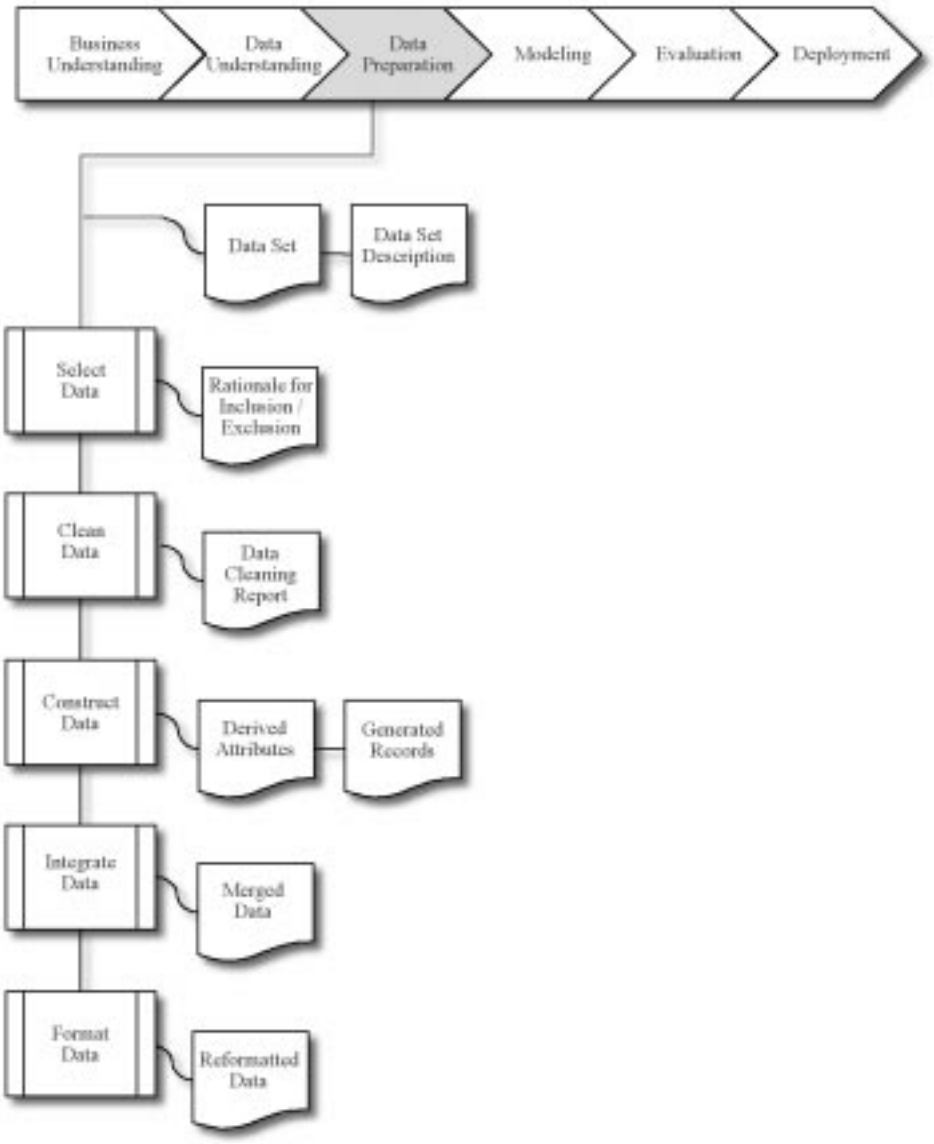


Figura 6: Preparación de datos

Salidas

conjunto de datos

Este es el conjunto de datos (o conjuntos de datos) producidos por la fase de preparación de datos, que se utilizará para el modelado o el trabajo de análisis principal del proyecto.

Descripción del conjunto de datos

Describa el conjunto de datos (o conjuntos de datos) que se utilizarán para el modelado o el trabajo de análisis principal del proyecto.

3.1 Seleccionar datos

Tarea **Seleccionar datos**

Decidir los datos que se utilizarán para el análisis. Los criterios incluyen la relevancia para los objetivos de minería de datos, la calidad y las restricciones técnicas, como los límites en el volumen de datos o los tipos de datos. Tenga en cuenta que la selección de datos cubre la selección de atributos (columnas) así como la selección de registros (filas) en una tabla.

Producción **Justificación de la inclusión/exclusión**

Enumere los datos que se incluirán/excluirán y las razones de estas decisiones.

3.2 Limpiar datos

Tarea **Limpiar datos**

Elevar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de subconjuntos limpios de datos, la inserción de valores predeterminados adecuados o técnicas más ambiciosas, como la estimación de datos faltantes mediante modelado.

Producción **Informe de limpieza de datos**

Describa qué decisiones y acciones se tomaron para abordar los problemas de calidad de los datos informados durante el*verificar la calidad de los datos*tarea de la *comprensión de datos*fase. Se deben considerar las transformaciones de los datos con fines de limpieza y el posible impacto en los resultados del análisis.

3.3 Construir datos

Tarea **Construir datos**

Esta tarea incluye operaciones de preparación de datos constructivos, como la producción de atributos derivados, registros nuevos completos o valores transformados para atributos existentes.

Salidas **Atributos derivados**

Los atributos derivados son atributos nuevos que se construyen a partir de uno o más atributos existentes en el mismo registro. Ejemplos:*área=largo * ancho.*

registros generados

Describir la creación de registros completamente nuevos. Ejemplo: crear registros para clientes que no realizaron ninguna compra durante el último año. No había ninguna razón para tener dichos registros en los datos sin procesar, pero para fines de modelado podría tener sentido representar explícitamente el hecho de que ciertos clientes no realizaron compras.

3.4 Integrar datos

Tarea

Integrar datos

Estos son métodos mediante los cuales la información se combina de varias tablas o registros para crear nuevos registros o valores.

Producción

datos combinados

Fusionar tablas se refiere a unir dos o más tablas que tienen información diferente sobre los mismos objetos. Ejemplo: una cadena minorista tiene una tabla con información sobre las características generales de cada tienda (p. ej., superficie, tipo de centro comercial), otra tabla con datos de ventas resumidos (p. ej., ganancias, cambio porcentual en las ventas del año anterior) y otra con información sobre la demografía de los alrededores. Cada una de estas tablas contiene un registro para cada tienda. Estas tablas se pueden fusionar en una nueva tabla con un registro para cada tienda, combinando campos de las tablas de origen.

Los datos combinados también cubren agregaciones. La agregación se refiere a operaciones en las que se calculan nuevos valores al resumir la información de múltiples registros y/o tablas. Por ejemplo, convertir una tabla de compras de clientes donde hay un registro para cada compra en una nueva tabla donde hay un registro para cada cliente, con campos como *número de compras, monto promedio de compra, porcentaje de pedidos cargados a tarjeta de crédito, porcentaje de artículos en promoción, etc.*

3.5 Formatear datos

Tarea

Formatear datos

Las transformaciones de formato se refieren principalmente a *sintáctico* modificaciones realizadas a los datos que no cambian su significado, pero que pueden ser requeridas por la herramienta de modelado.

Producción

Datos reformateados

Algunas herramientas tienen requisitos sobre el orden de los atributos, como que el primer campo sea un identificador único para cada registro o que el último campo sea el campo de resultado que el modelo debe predecir.

Puede ser importante cambiar el orden de los registros en el conjunto de datos. Tal vez la herramienta de modelado requiera que los registros se clasifiquen según el valor del atributo de resultado. Una situación común es que los registros del conjunto de datos se ordenen inicialmente de alguna manera, pero el algoritmo de modelado necesita que estén en un orden bastante aleatorio. Por ejemplo, cuando se usan redes neuronales, generalmente es mejor que los registros se presenten en un orden aleatorio, aunque algunas herramientas manejan esto automáticamente sin la intervención explícita del usuario.

Además, se realizan cambios puramente sintácticos para satisfacer los requisitos de la herramienta de modelado específica. Ejemplos: eliminar comas de los campos de texto en archivos de datos delimitados por comas, recortar todos los valores a un máximo de 32 caracteres.

4 Modelado

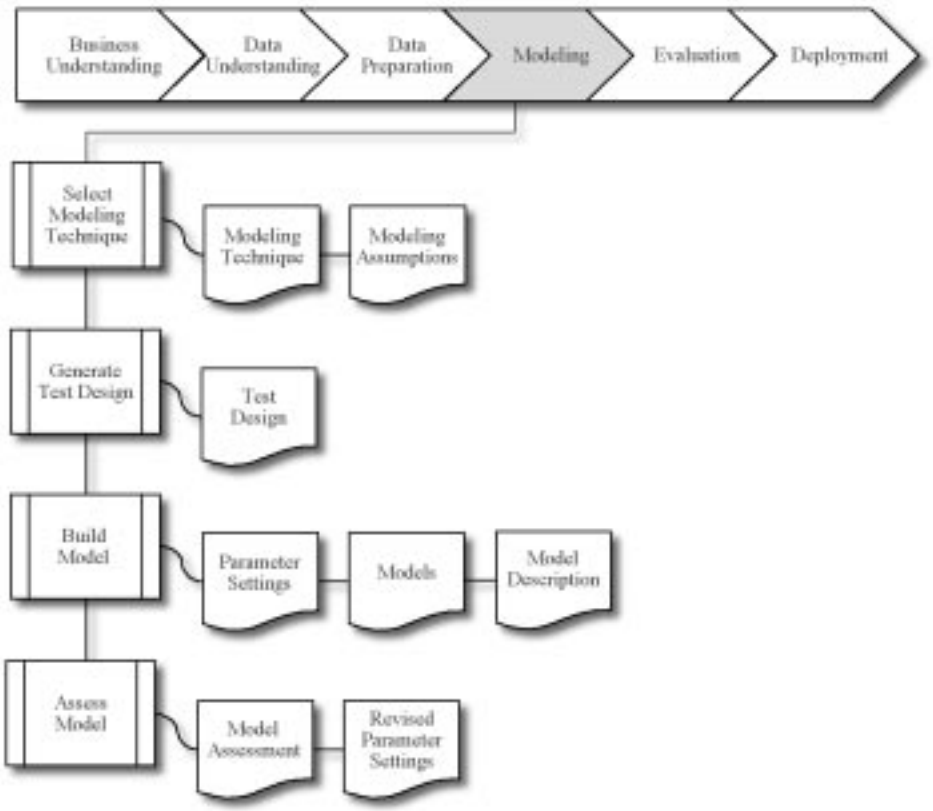


Figura 7: Modelado

4.1 Seleccionar técnica de modelado

Tarea **Seleccionar técnica de modelado**

Como primer paso en el modelado, seleccione la técnica de modelado real que se va a utilizar. Si bien es posible que ya haya seleccionado una herramienta de comprensión empresarial, esta tarea se refiere a la técnica de modelado específica, por ejemplo, la creación de árboles de decisión con C4.5 o la generación de redes neuronales con propagación hacia atrás. Si se aplican varias técnicas, realice esta tarea para cada técnica por separado.

Salidas **Técnica de modelado**

Documente la técnica de modelado real que se va a utilizar.

Supuestos de modelado

Muchas técnicas de modelado hacen suposiciones específicas sobre los datos, por ejemplo, todos los atributos tienen distribuciones uniformes, no se permiten valores faltantes, el atributo de clase debe ser simbólico, etc. Registre cualquier suposición hecha.

4.2 Generar diseño de prueba

Tarea

Generar diseño de prueba

Antes de que realmente construyamos un modelo, necesitamos generar un procedimiento o mecanismo para probar la calidad y validez del modelo. Por ejemplo, en tareas de minería de datos supervisadas como la clasificación, es común utilizar tasas de error como medidas de calidad para los modelos de minería de datos. Por lo tanto, normalmente separamos el conjunto de datos en tren y conjunto de prueba, construimos el modelo en el conjunto de tren y estimamos su calidad en el conjunto de prueba separado.

Producción

Diseño de prueba

Describa el plan previsto para entrenar, probar y evaluar los modelos. Un componente principal del plan es decidir cómo dividir el conjunto de datos disponible en datos de entrenamiento, datos de prueba y conjuntos de datos de validación.

4.3 Modelo de construcción

Tarea

Construir modelo

Ejecute la herramienta de modelado en el conjunto de datos preparado para crear uno o más modelos.

Salidas

Configuración de parámetros

Con cualquier herramienta de modelado, a menudo hay una gran cantidad de parámetros que se pueden ajustar. Enumere los parámetros y su valor elegido, junto con la justificación de la elección de la configuración de los parámetros.

Modelos

Estos son los modelos reales producidos por la herramienta de modelado, no un informe.

Descripcion del modelo

Describe el modelo resultante. Informar sobre la interpretación de los modelos y documentar las dificultades encontradas con sus significados.

4.4 Modelo de evaluación

Tarea

Evaluar modelo

El ingeniero de minería de datos interpreta los modelos de acuerdo con su conocimiento del dominio, los criterios de éxito de la minería de datos y el diseño de prueba deseado. Esta tarea interfiere con la fase de evaluación posterior. Mientras que el ingeniero de minería de datos juzga el éxito de la aplicación de técnicas de modelado y descubrimiento de manera más técnica, luego se pone en contacto con analistas comerciales y expertos en el dominio para analizar los resultados de la minería de datos en el contexto comercial. Además, esta tarea solo considera modelos, mientras que la fase de evaluación también tiene en cuenta todos los demás resultados que se produjeron en el transcurso del proyecto.

El ingeniero de minería de datos intenta clasificar los modelos. Evalúa los modelos según los criterios de evaluación. En la medida de lo posible, también tiene en cuenta los objetivos empresariales y los criterios de éxito empresarial. En la mayoría de los proyectos de minería de datos, el ingeniero de minería de datos aplica una sola técnica más de una vez o genera resultados de minería de datos con diferentes técnicas alternativas. En esta tarea, también compara todos los resultados según los criterios de evaluación.

Salidas

Evaluación del modelo

Resuma los resultados de esta tarea, enumere las cualidades de los modelos generados (p. ej., en términos de precisión) y clasifique su calidad en relación con los demás.

Ajustes de parámetros revisados

De acuerdo con la evaluación del modelo, revise la configuración de los parámetros y ajústelos para la próxima ejecución en la tarea Construir modelo. Repita la construcción y evaluación del modelo hasta que crea firmemente que encontró el *mejor* modelo(s). Documente todas esas revisiones y evaluaciones.

5 Evaluación

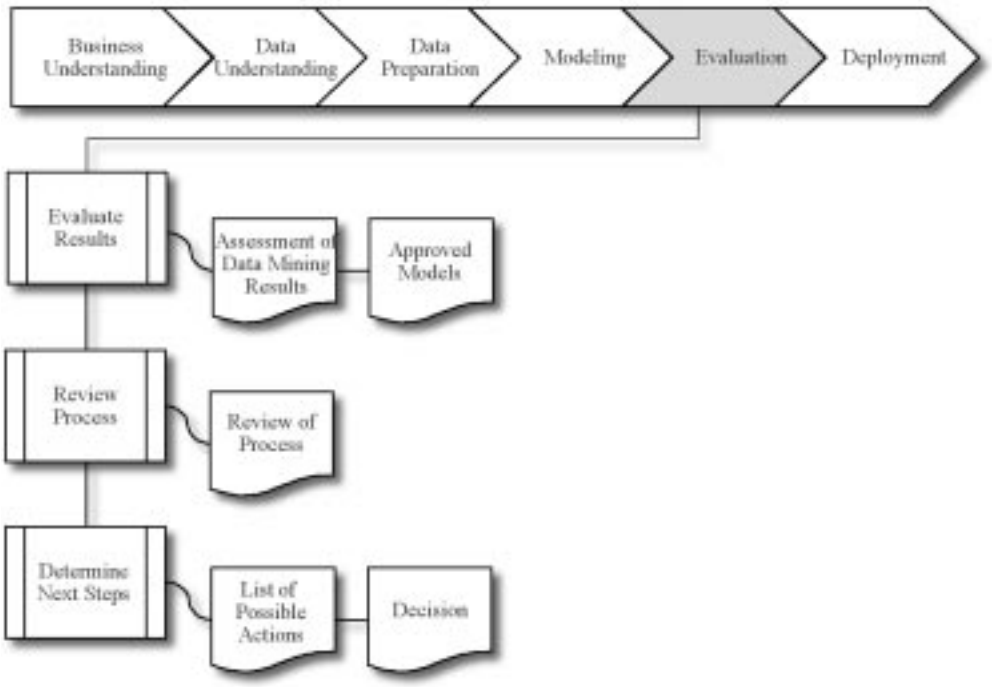


Figura 8: Evaluación

5.1 Evaluar resultados

Tarea

Evaluar resultados

Los pasos de evaluación anteriores se ocuparon de factores como la precisión y la generalidad del modelo. Este paso evalúa el grado en que el modelo cumple con los objetivos comerciales y busca determinar si existe alguna razón comercial por la cual este modelo es deficiente. Otra opción de evaluación es probar los modelos en aplicaciones de prueba en la aplicación real si las limitaciones de tiempo y presupuesto lo permiten.

Además, la evaluación también evalúa otros resultados de minería de datos generados. Los resultados de la minería de datos cubren modelos que están necesariamente relacionados con los objetivos comerciales originales y todos los demás hallazgos que no están necesariamente relacionados con los objetivos comerciales originales, pero que también pueden revelar desafíos adicionales, información o sugerencias para direcciones futuras.

Salidas

Evaluación de los resultados de la minería de datos con respecto a los criterios de éxito empresarial

Resuma los resultados de la evaluación en términos de criterios de éxito comercial, incluida una declaración final sobre si el proyecto ya cumple con los objetivos comerciales iniciales.

Modelos homologados

Después de la evaluación del modelo con respecto a los criterios de éxito comercial, los modelos generados que cumplen con los criterios seleccionados se convierten en modelos aprobados.

5.2 Proceso de revisión

Tarea

Proceso de revisión

En este punto, se espera que el modelo resultante parezca satisfactorio y satisfaga las necesidades comerciales. Ahora es apropiado hacer una revisión más exhaustiva del compromiso de minería de datos para determinar si hay algún factor o tarea importante que de alguna manera se haya pasado por alto. Esta revisión también cubre cuestiones de garantía de calidad, por ejemplo, ¿construimos correctamente el modelo? ¿Solo usamos atributos que tenemos permitido usar y que están disponibles para análisis futuros?

Producción

Revisión del proceso

Resuma la revisión del proceso y resalte las actividades que se han perdido y/o deben repetirse.

5.3 Determinar los próximos pasos

Tarea

Determinar los próximos pasos

De acuerdo con los resultados de la evaluación y la revisión del proceso, el proyecto decide cómo proceder en esta etapa. El proyecto debe decidir si finalizar este proyecto y pasar a la implementación si corresponde o si iniciar más iteraciones o configurar nuevos proyectos de minería de datos. Esta tarea incluye análisis de recursos remanentes y presupuesto que influye en las decisiones.

Salidas

Lista de posibles acciones

Enumere las posibles acciones adicionales junto con las razones a favor y en contra de cada opción.

Decisión

Describa la decisión sobre cómo proceder junto con la justificación.

6 Despliegue

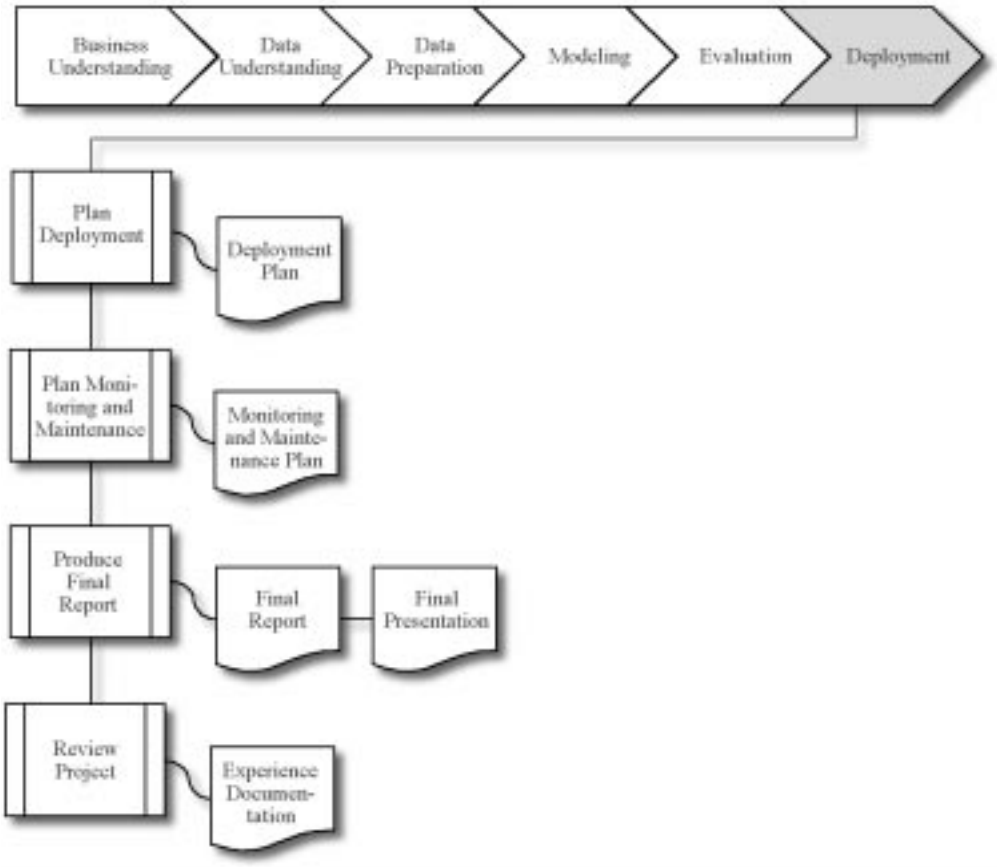


Figura 9: Implementación

6.1 Planificar el despliegue

Tarea

Planificar la implementación

Para implementar los resultados de la minería de datos en el negocio, esta tarea toma los resultados de la evaluación y concluye una estrategia para la implementación.
Si se ha identificado un procedimiento general para crear los modelos relevantes, este procedimiento se documenta aquí para una implementación posterior.

Producción

Plan de empleo

Resuma la estrategia de implementación, incluidos los pasos necesarios y cómo realizarlos.

6.2 Seguimiento y mantenimiento del plan

Tarea

Monitoreo y mantenimiento del plan

El monitoreo y el mantenimiento son temas importantes si el resultado de la minería de datos se convierte en parte del día a día del negocio y su entorno. Una preparación cuidadosa de una estrategia de mantenimiento ayuda a evitar períodos innecesariamente largos de uso incorrecto de los resultados de la minería de datos. Para monitorear el despliegue de los resultados de la minería de datos, el proyecto necesita un plan detallado sobre el proceso de monitoreo. Este plan tiene en cuenta el tipo específico de implementación.

Producción

Plan de seguimiento y mantenimiento

Resumir la estrategia de monitoreo y mantenimiento, incluidos los pasos necesarios y cómo realizarlos.

6.3 Producir informe final

Tarea

Producir informe final

Al final del proyecto, el líder del proyecto y su equipo redactan un informe final. Según el plan de implementación, este informe puede ser solo un resumen del proyecto y sus experiencias (si aún no se han documentado como una actividad en curso) o puede ser una presentación final y completa de los resultados de la extracción de datos.

Salidas

Reporte final

Este es el informe escrito final del compromiso de minería de datos. Incluye todos los entregables anteriores y resume y organiza los resultados.

Presentación final

A menudo también habrá una reunión al final del proyecto donde los resultados se presentan verbalmente al cliente.

6.4 Proyecto de revisión

Tarea

Revisar proyecto

Evaluar lo que salió bien y lo que salió mal, lo que se hizo bien y lo que debe mejorarse.

Producción

Documentación de la experiencia

Resuma las experiencias importantes realizadas durante el proyecto. Por ejemplo, las trampas, los enfoques engañosos o las sugerencias para seleccionar las técnicas de minería de datos más adecuadas en situaciones similares podrían ser parte de esta documentación. En proyectos ideales, la documentación de la experiencia cubre también cualquier informe que haya sido escrito por miembros individuales del proyecto durante las fases del proyecto y sus tareas.

III La guía del usuario de CRISP-DM

1 Comprensión empresarial

1.1 Determinar los objetivos comerciales

Tarea

Determinar los objetivos comerciales.

El primer objetivo del analista es comprender a fondo, desde una*negocio* perspectiva, lo que el cliente realmente quiere lograr. A menudo, el cliente tiene muchos objetivos y limitaciones en competencia que deben equilibrarse adecuadamente. El objetivo del analista es descubrir factores importantes al comienzo del proyecto que pueden influir en el resultado final. Una consecuencia probable de descuidar este paso sería dedicar una gran cantidad de esfuerzo a producir las respuestas correctas a las preguntas incorrectas.

Producción

Fondo

Cotejar la información que se conoce sobre la situación empresarial de la organización al inicio del proyecto. Estos detalles no solo sirven para identificar más de cerca los objetivos comerciales que se resolverán, sino que también sirven para identificar los recursos, tanto humanos como materiales, que pueden usarse o necesitarse durante el transcurso del proyecto.

Actividades

Organización

- Desarrollar organigramas que identifiquen divisiones, departamentos y grupos de proyectos. El cuadro también debe identificar los nombres y responsabilidades de los gerentes.
- Identificar las personas clave en el negocio y sus roles.
- Identifique un patrocinador interno (patrocinador financiero y usuario principal/experto en el dominio).
- ¿Existe un comité directivo y quiénes son los miembros?
- Identifique las unidades de negocio que se ven afectadas por el proyecto de minería de datos (p. ej., marketing, ventas, finanzas).

área problemática

- Identifique el área problemática (p. ej., marketing, atención al cliente, desarrollo comercial, etc.)**.**
- Describe el problema en términos generales.
- Comprobar el estado actual del proyecto (p. ej., comprobar si ya está claro dentro de la unidad de negocio que estamos realizando un proyecto de minería de datos o necesitamos anunciar la minería de datos como una tecnología clave en el negocio?).
- Aclare los requisitos previos del proyecto (p. ej., ¿cuál es la motivación del proyecto? ¿La empresa ya utiliza minería de datos?).
- Si es necesario, preparar presentaciones y presentar data mining al negocio.

- Identifique los grupos objetivo para el resultado del proyecto (p. ej., ¿esperamos un informe escrito para la alta dirección o esperamos un sistema en ejecución que sea utilizado por usuarios finales ingenuos?).
- Identificar las necesidades y expectativas de los usuarios.

Solución actual

- Describa cualquier solución actualmente en uso para el problema.
- Describa las ventajas y desventajas de la solución actual y el nivel de aceptación de los usuarios.

Producción

Objetivos de negocios

Describa el objetivo principal del cliente, desde una perspectiva comercial, en el proyecto de minería de datos. Además del objetivo comercial principal, normalmente hay una gran cantidad de preguntas comerciales relacionadas que el cliente desea abordar. Por ejemplo, el objetivo comercial principal podría ser mantener a los clientes actuales prediciendo cuándo son propensos a cambiarse a un competidor, mientras que los objetivos comerciales secundarios podrían ser determinar si las tarifas más bajas afectan solo a un segmento particular de clientes.

Actividades

- Describa informalmente el problema que se supone que debe resolverse con la minería de datos.
- Especifique todas las preguntas comerciales con la mayor precisión posible.
- Especifique cualquier otro requisito comercial (p. ej., la empresa no quiere perder ningún cliente).
- Especifique los beneficios esperados en términos comerciales.

¡Tener cuidado!

Tenga cuidado con establecer objetivos inalcanzables: hágalos lo más realistas posible.

Producción

Criterios de éxito empresarial

Describa los criterios para un resultado exitoso o útil para el proyecto desde el punto de vista empresarial. Esto puede ser bastante específico y fácilmente medible, como la reducción de la rotación de clientes a un cierto nivel o general y subjetivo, como "brindar información útil sobre las relaciones". En este último caso se debe indicar quién haría el juicio subjetivo.

Actividades

- Especificar criterios de éxito empresarial (por ejemplo, mejorar la tasa de respuesta en una campaña de correo en un 10 por ciento y aumentar la tasa de registro en un 20 por ciento).
- Identificar quién evalúa los criterios de éxito.

¡Recordar!

Cada uno de los criterios de éxito debe relacionarse con *al menos uno* de los objetivos comerciales especificados.

¡Buena idea! Antes de comenzar con la evaluación de la situación, puede considerar experiencias previas de este problema, ya sea internamente usando CRISP-DM o externamente usando soluciones preempaquetadas.

1.2 Evaluar la situación

Tarea **evaluar la situación**

Esta tarea implica una búsqueda de hechos más detallada sobre todos los recursos, restricciones, suposiciones y otros factores que se deben considerar para determinar el objetivo del análisis de datos y el plan del proyecto.

Producción

Inventario de recursos

Enumere los recursos disponibles para el proyecto, incluidos: personal (expertos en negocios y datos, soporte técnico, personal de minería de datos), datos (extractos fijos, acceso a datos operativos o almacenados en vivo), recursos informáticos (plataformas de hardware), software (minería de datos herramientas, otro software relevante).

Actividades

Recursos de hardware

- Identifique el hardware base.
- Establecer la disponibilidad del hardware base para el proyecto de minería de datos.
- Compruebe si el programa de mantenimiento del hardware entra en conflicto con la disponibilidad del hardware para el proyecto de minería de datos.
- Identifique el hardware disponible para la herramienta de minería de datos que se utilizará (si la herramienta se conoce en esta etapa).

Fuentes de datos y conocimiento

- Identificar las fuentes de datos.
- Identificar el tipo de fuentes de datos (fuentes en línea, expertos, documentación escrita, etc.).
- Identificar fuentes de conocimiento.
- Identificar el tipo de fuentes de conocimiento (fuentes en línea, expertos, documentación escrita, etc.).
- Consultar herramientas y técnicas disponibles.
- Describa el conocimiento previo relevante (informal o formal).

Fuentes de personal

- Identificar el patrocinador del proyecto (si es diferente del patrocinador interno como en la Sección 1.1.1).
- Identifique al administrador del sistema, al administrador de la base de datos y al personal de soporte técnico para más preguntas.
- Identifique analistas de mercado, expertos en minería de datos y estadísticos y verifique su disponibilidad.
- Consultar disponibilidad de expertos de dominio para fases posteriores.

¡Recordar! Recuerde que el proyecto puede necesitar personal técnico en momentos extraños a lo largo del proyecto, por ejemplo, durante la Transformación de datos.

Producción

Requisitos, suposiciones y restricciones

Enumere todos los requisitos del proyecto, incluido el calendario de finalización, la comprensibilidad y la calidad de los resultados y la seguridad, así como las cuestiones legales. Como parte de este resultado, asegúrese de que tiene permiso para usar los datos.

Enumere las suposiciones hechas por el proyecto. Estas pueden ser suposiciones sobre los datos, que se pueden verificar durante la extracción de datos, pero también pueden incluir suposiciones no verificables sobre el negocio en el que se basa el proyecto. Es particularmente importante enumerar estos últimos si forman condiciones sobre la validez de los resultados.

Enumere las restricciones impuestas al proyecto. Estas restricciones pueden implicar la falta de recursos para llevar a cabo algunas de las tareas del proyecto dentro del plazo requerido o puede haber restricciones legales o éticas sobre el uso de los datos o la solución necesaria para llevar a cabo la tarea de minería de datos.

Actividades

Requisitos

- Especifique el perfil del grupo objetivo.
- Capture todos los requisitos en la programación.
- Capture los requisitos de comprensibilidad, precisión, capacidad de implementación, mantenibilidad y repetibilidad del proyecto de minería de datos y los modelos resultantes.
- Capture los requisitos sobre seguridad, restricciones legales, privacidad, informes y cronograma del proyecto.

suposiciones

- Aclarar todos los supuestos (incluidos los implícitos) y hacerlos explícitos (p. ej., para abordar la cuestión empresarial, es necesario un número mínimo de clientes con una edad superior a 50 años).
- Enumere los supuestos sobre la calidad de los datos (p. ej., precisión, disponibilidad).
- Haga una lista de supuestos sobre factores externos (p. ej., problemas económicos, productos competitivos, avances técnicos).
- Aclare las suposiciones que conducen a cualquiera de las estimaciones (p. ej., se supone que el precio de una herramienta específica es inferior a $1000).
- Enumere todas las suposiciones sobre si es necesario *entender* y describir o explicar el modelo. (Por ejemplo, cómo se deben presentar el modelo y los resultados a la alta gerencia/patrocinador).

Restricciones

- Verifique las limitaciones generales (p. ej., cuestiones legales, presupuesto, plazos y recursos).
- Compruebe los derechos de acceso a las fuentes de datos (p. ej., restricciones de acceso, se requiere contraseña).
- Comprobar la accesibilidad técnica de los datos (sistemas operativos, sistema de gestión de datos, formato de archivo o base de datos).
- Compruebe si el conocimiento relevante es accesible.
- Verificar restricciones presupuestarias (Costos fijos, costos de implementación, etc.).

¡Recordar! La lista de supuestos también incluye supuestos al inicio del proyecto, es decir, cuál ha sido el punto de partida del proyecto.

Producción

Riesgos y contingencias

Enumere los riesgos, es decir, los eventos que pueden ocurrir, impactando el cronograma, el costo o el resultado. Listar los planes de contingencia correspondientes; qué acción se tomará para evitar o minimizar el impacto o recuperarse de la ocurrencia de los riesgos previstos.

Actividades

Identificar riesgos

- Identifique los riesgos comerciales (p. ej., el competidor obtiene mejores resultados primero).
- Identifique los riesgos organizacionales (p. ej., el departamento que solicita el proyecto no tiene fondos para el proyecto).
- Identificar los riesgos financieros (p. ej., la financiación adicional depende de los resultados iniciales de la extracción de datos).
- Identificar riesgos técnicos.
- Identificar los riesgos que dependen de los datos y las fuentes de datos (p. ej., mala calidad y cobertura).

Desarrollar planes de contingencia.

- Determinar las condiciones bajo las cuales puede ocurrir cada riesgo.
- Desarrollar planes de contingencia.

Producción

Terminología

Compilar un glosario de terminología relevante para el proyecto. Esto debe incluir al menos dos componentes:

(1) Un glosario de terminología comercial relevante, que forma parte de la comprensión comercial disponible para el proyecto.

(2) Un glosario de terminología de minería de datos, ilustrado con ejemplos relevantes para el problema comercial en cuestión.

Actividades

- Consultar previa disponibilidad de glosarios, caso contrario comenzar a redactar glosarios.
- Hable con expertos en dominios para comprender su terminología.
- Familiarizarse con la terminología empresarial.

Producción

Costos y beneficios

Prepare un análisis de costo-beneficio para el proyecto, comparando los costos del proyecto con el beneficio potencial para el negocio si tiene éxito.

¡Buena idea!

La comparación debe ser lo más específica posible, ya que esto permite hacer un mejor caso de negocios.

Actividades

- Estimar los costos de la recopilación de datos.
- Estimar los costos de desarrollar e implementar una solución.
- Identifique los beneficios cuando se implementa una solución (por ejemplo, mayor satisfacción del cliente, ROI y aumento de los ingresos).
- Estimar los costos de operación.

¡Tener cuidado!

Recuerde identificar los costos ocultos, como la extracción y preparación repetida de datos, los cambios en los flujos de trabajo y el tiempo de capacitación durante el aprendizaje.

1.3 Determinar los objetivos de la minería de datos

Tarea

Determinar los objetivos de minería de datos

A*objetivo de negocio*establece objetivos en terminología empresarial; a*objetivo de minería de datos*establece los objetivos del proyecto en términos técnicos. Por ejemplo, el objetivo comercial podría ser "Aumentar las ventas por catálogo a los clientes existentes", mientras que un objetivo de minería de datos podría ser "Predecir cuántos widgets comprará un cliente, dadas sus compras en los últimos tres años, la información demográfica relevante y el precio del producto". artículo."

Producción

Objetivos de minería de datos

Describa los resultados previstos del proyecto que permiten el logro de los objetivos comerciales. Tenga en cuenta que estos son normalmente*técnico*salidas.

Actividades

- Traducir las preguntas comerciales a objetivos de minería de datos (p. ej., una campaña de marketing requiere la segmentación de clientes para decidir a quién abordar en esta campaña; se debe especificar el nivel/tamaño de los segmentos).
- Especifique el tipo de problema de minería de datos (p. ej., clasificación, descripción, predicción y agrupación). Para obtener más detalles sobre los tipos de problemas de minería de datos, consulte el Apéndice V.2, donde se describen con más detalle.

¡Buena idea! Puede ser conveniente redefinir el problema. Por ejemplo, modelar la retención de productos en lugar de la retención de clientes, ya que enfocarse en la retención de clientes puede ser demasiado tarde para afectar el resultado.

Producción **Criterios de éxito de la minería de datos**

Defina los criterios para un resultado exitoso del proyecto en términos técnicos, por ejemplo, un cierto nivel de precisión predictiva o un perfil de propensión a comprar con un grado determinado de "elevación". Al igual que con los criterios de éxito comercial, puede ser necesario describirlos en términos subjetivos, en cuyo caso se debe identificar a la persona o personas que realizan el juicio subjetivo.

Actividades

- Especificar criterios para la evaluación del modelo (p. ej., precisión, rendimiento y complejidad del modelo).
- Definir puntos de referencia para los criterios de evaluación.
- Especificar los criterios que abordan los criterios de evaluación subjetivos (por ejemplo, la capacidad de explicación del modelo y los datos y la percepción de marketing proporcionados por el modelo).

¡Tener cuidado! Recuerde que los criterios de éxito de la minería de datos son diferentes a los criterios de éxito empresarial definidos anteriormente.

Recuerde que es aconsejable planificar la implementación desde el inicio del proyecto.

1.4 Producir plan de proyecto

Tarea **Producir el plan de proyecto**

Describa el plan previsto para lograr los objetivos de minería de datos y, por lo tanto, lograr los objetivos comerciales.

Producción **Plan de proyecto**

Enumere las etapas a ejecutar en el proyecto, junto con la duración, los recursos requeridos, las entradas, las salidas y las dependencias. Siempre que sea posible, haga explícitas las iteraciones a gran escala en el proceso de extracción de datos, por ejemplo, repeticiones de las fases de modelado y evaluación. Como parte del plan del proyecto, también es importante analizar las dependencias entre el cronograma y los riesgos. Marque los resultados de estos análisis explícitamente en el plan del proyecto, idealmente con acciones y recomendaciones para acciones si aparecen los riesgos.

Recuerde que, aunque esta es la única tarea en la que se nombra directamente el Plan del Proyecto, sin embargo, debe consultarse y revisarse continuamente a lo largo del proyecto. Al menos debe consultarse cada vez que se inicia una nueva tarea o se inicia una nueva iteración de una tarea o actividad.

Actividades

- Defina el plan de proceso inicial y discuta la factibilidad con todo el personal involucrado.
- Reúna todos los objetivos identificados y las técnicas seleccionadas en un procedimiento coherente que resuelva las preguntas comerciales y cumpla con los criterios de éxito empresarial.
- Estimar el esfuerzo y los recursos necesarios para lograr e implementar la solución (es útil considerar la experiencia de otras personas al estimar escalas de tiempo para proyectos de minería de datos. Por ejemplo, a menudo se postula que el 50-70 por ciento del tiempo y esfuerzo en un proyecto de minería de datos se utiliza en la Fase de preparación de datos y entre un 20 y un 30 por ciento en la Fase de comprensión de datos, mientras que solo entre un 10 y un 20 por ciento se gasta en cada una de las Fases de modelado, evaluación y comprensión empresarial y entre un 5 y un 10 por ciento en la Fase de implementación).
- Identificar los pasos críticos.
- Marcar puntos de decisión.
- Marcar puntos de revisión.
- Identificar las principales iteraciones.

Producción

Evaluación inicial de herramientas y técnicas.

Al final de la primera fase, el proyecto también realiza una evaluación inicial de herramientas y técnicas. Aquí, selecciona una herramienta de minería de datos que admita varios métodos para diferentes etapas del proceso, por ejemplo. Es importante evaluar las herramientas y técnicas al principio del proceso, ya que la selección de herramientas y técnicas posiblemente influya en todo el proyecto.

Actividades

- Cree una lista de criterios de selección para herramientas y técnicas (o use una existente si está disponible).
- Elija herramientas y técnicas potenciales.
- Evaluar la adecuación de las técnicas.
- Revisar y priorizar técnicas aplicables de acuerdo a la evaluación de alternativas de solución.

2 Comprensión de datos

2.1 Recolectar datos iniciales

Tarea

Recopilar datos iniciales

Adquirir dentro del proyecto los datos (o acceder a los datos) enumerados en los recursos del proyecto. Esta recopilación inicial incluye la carga de datos si es necesario para la comprensión de los datos. Por ejemplo, si tiene la intención de utilizar una herramienta específica para la comprensión de datos, es lógico cargar sus datos en esta herramienta.

Producción

Informe inicial de recopilación de datos

Enumere todos los diversos datos utilizados en el proyecto, junto con los requisitos de selección para obtener datos más detallados. El informe de recopilación de datos también debe definir si algunos atributos son relativamente más importantes que otros.

Recuerde que cualquier evaluación de la calidad de los datos debe hacerse no solo de las fuentes de datos individuales, sino también de cualquier dato que provenga de la fusión de fuentes de datos. Los datos combinados pueden presentar problemas que no existen en las fuentes de datos individuales debido a las incoherencias entre las fuentes.

Actividades

Planificación de requisitos de datos

- Planifique qué información se necesita (por ejemplo, solo atributos dados, información adicional).
- Compruebe si toda la información necesaria (para resolver los objetivos de minería de datos) está realmente disponible.

Criteria de selección

- Especifique los criterios de selección (p. ej., ¿qué atributos son necesarios para los objetivos de minería de datos especificados? ¿Qué atributos se han identificado como irrelevantes? ¿Cuántos atributos podemos manejar con las técnicas elegidas?).
- Seleccionar tablas/archivos de interés.
- Seleccionar datos dentro de una tabla/archivo.
- Piense en cuánto tiempo se debe usar el historial, incluso si está disponible (por ejemplo, incluso si se dispone de datos de 18 meses, quizás solo se necesiten 12 meses para el ejercicio).

¡Tener cuidado!

Tenga en cuenta que los datos recopilados de diferentes fuentes pueden dar lugar a problemas de calidad cuando se combinan (por ejemplo, los archivos de direcciones combinados con la propia base de clientes pueden mostrar inconsistencias de formato, datos no válidos, etc.).

Inserción de datos

- Si los datos contienen entradas de texto libre, ¿necesitamos codificarlos para el modelado o queremos agrupar entradas específicas?
- ¿Cómo se pueden adquirir los atributos que faltan?
- Describe cómo extraer los datos.

¡Buena idea! Recuerde que parte del conocimiento sobre los datos puede estar en fuentes no electrónicas (por ejemplo, Personas, Texto impreso, etc.).

Recuerde que puede ser necesario preprocesar los datos (datos de series temporales, promedios ponderados, etc.).

2.2 Describir datos

Tarea **Describir datos**

Examine las propiedades "brutas" de los datos adquiridos e informe sobre los resultados.

Producción **Informe de descripción de datos**

Describa los datos que se han adquirido, incluidos: el formato de los datos, la cantidad de datos (por ejemplo, el número de registros y campos dentro de cada tabla), las identidades de los campos y cualquier otra característica superficial de los datos que se han descubierto. .

Actividades **Análisis volumétrico de datos**

- Identificar datos y método de captura.
- Acceder a fuentes de datos.
- Use análisis estadísticos si es apropiado.
- Tablas de informes y sus relaciones.
- Verifique el volumen de datos, el número de múltiplos, la complejidad.
- ¿Los datos contienen entradas de texto libre?

Tipos y valores de atributos

- Verifique la accesibilidad y disponibilidad de los atributos.
- Compruebe los tipos de atributos (numéricos, simbólicos, taxonómicos, etc.).
- Verifique los rangos de valores de los atributos.
- Analizar las correlaciones de atributos.
- Comprender el significado de cada atributo y valor de atributo en términos comerciales.
- Para cada atributo, calcule estadísticas básicas (p. ej., calcule distribución, promedio, máximo, mínimo, desviación estándar, varianza, moda, asimetría, etc.).
- Analice estadísticas básicas y relacione los resultados con su significado en términos comerciales.
- ¿El atributo es relevante para el objetivo específico de minería de datos?

- ¿Se usa consistentemente el significado del atributo?
- Entreviste a un experto en el dominio sobre su opinión sobre la relevancia de los atributos.
- ¿Es necesario balancear los datos (Dependiendo de la técnica de modelado utilizada)?

Llaves

- Analizar las relaciones clave.
- Verifique la cantidad de superposiciones de valores de atributos clave en las tablas.

Revisar suposiciones/objetivos

- Actualice la lista de suposiciones si es necesario.

2.3 Explorar datos

Tarea

Explorar datos

Esta tarea aborda las preguntas de minería de datos que se pueden abordar mediante consultas, visualización e informes. Estos análisis pueden abordar directamente los objetivos de minería de datos. Sin embargo, también pueden contribuir o refinar la descripción de datos y los informes de calidad y alimentar la transformación y otra preparación de datos necesaria para un análisis posterior.

Producción

Informe de exploración de datos

Describa los resultados de esta tarea, incluidos los primeros hallazgos o las hipótesis iniciales y su impacto en el resto del proyecto. El informe también puede cubrir gráficos y diagramas que indican las características de los datos o conducen a subconjuntos de datos interesantes para un examen más detallado.

Actividades

Exploración de datos

- Analizar las propiedades de los atributos interesantes en detalle (p. ej., estadísticas básicas de subpoblaciones interesantes).
- Identificar las características de las subpoblaciones.

Formar suposiciones para futuros análisis.

- Considere y evalúe la información y los hallazgos en el informe de descripciones de datos.
- Formular hipótesis e identificar acciones.
- Transforme la hipótesis en un objetivo de minería de datos si es posible.
- Aclare los objetivos de minería de datos o hágalos más precisos. La búsqueda a ciegas no es necesariamente inútil, pero es preferible una búsqueda más dirigida hacia los objetivos comerciales.
- Realizar análisis básicos para verificar la hipótesis.

2.4 Verificar la calidad de los datos

Tarea

Verificar la calidad de los datos

Examinar la calidad de los datos, abordando preguntas como: ¿Están completos los datos (cubren todos los casos requeridos)? ¿Es correcto o contiene errores y, si hay errores, qué tan comunes son? ¿Hay valores faltantes en los datos? Si es así, ¿cómo se representan, dónde ocurren y qué tan comunes son?

Producción

Informe de calidad de datos

Listar los resultados de la verificación de la calidad de los datos; si hay problemas de calidad, enumere las posibles soluciones.

Actividades

- Identificar valores especiales y catalogar su significado.

Revisar claves, atributos

- Compruebe la cobertura (por ejemplo, están representados todos los valores posibles).
- Comprobar llaves.
- ¿Coinciden los significados de los atributos y los valores contenidos?
- Identifique los atributos que faltan y los campos en blanco.
- Significado de los datos faltantes.
- Busque atributos con diferentes valores que tengan significados similares (p. ej., dieta baja en grasas).
- Verifique la ortografía de los valores (p. ej., el mismo valor pero a veces comienza con una letra minúscula, a veces con una letra mayúscula).
- Compruebe si hay desviaciones, decida si una desviación es ruido o puede indicar un fenómeno interesante.
- Compruebe la verosimilitud de los valores, por ejemplo, todos los campos tienen los mismos o casi los mismos valores.

¡Buena idea!

Haga una revisión de cualquier atributo que pueda dar respuestas que entren en conflicto con el sentido común (por ejemplo, adolescentes con altos ingresos)

Use gráficos de visualización, histogramas, etc. para mostrar inconsistencias en los datos

Calidad de datos en archivos planos

- Si los datos se almacenan en archivos sin formato, verifique qué delimitador se usa y si se usa de manera consistente dentro de todos los atributos.
- Si los datos se almacenan en archivos planos, verifique la cantidad de campos en cada registro. ¿Coinciden?

Ruido e inconsistencias entre fuentes.

- Comprobar consistencias y redundancias entre diferentes fuentes.
- Planifique cómo lidiar con el ruido.
- Detecta el tipo de ruido y qué atributos se ven afectados.

¡Buena idea! Recuerde que puede ser necesario excluir algunos datos ya que no muestran un comportamiento ni positivo ni negativo (por ejemplo, para verificar el comportamiento crediticio de los clientes, excluir a todos aquellos que nunca han prestado, no financian una hipoteca de vivienda, aquellos cuya hipoteca está a punto de vencimiento , etc.

Revise todas las suposiciones, ya sean válidas o no, dada la información actual sobre datos y conocimientos.

3 Preparación de datos

Producción **conjunto de datos**

Este es el conjunto de datos (o conjuntos de datos) producidos por la fase de preparación de datos, utilizados para el modelado o el trabajo de análisis principal del proyecto.

Producción **Descripción del conjunto de datos**

Esta es la descripción de los conjuntos de datos utilizados para el modelado o el trabajo de análisis principal del proyecto.

3.1 Seleccionar datos

Tarea **Seleccionar datos**

Decidir los datos que se utilizarán para el análisis. Los criterios incluyen la relevancia para los objetivos de minería de datos, la calidad y las restricciones técnicas, como los límites en el volumen de datos o los tipos de datos.

Producción **Justificación de la inclusión/exclusión**

Enumere los datos que se usarán/excluirán y las razones de estas decisiones.

Actividades

- Recopile datos adicionales apropiados (de diferentes fuentes, tanto internas como externas).
- Realice pruebas de significación y correlación para decidir si se deben incluir campos.
- Reconsidere los criterios de selección de datos (consulte la Tarea 2.1) a la luz de las experiencias de calidad de datos, exploración de datos (es decir, puede desear incluir/excluir otros conjuntos de datos).
- Reconsidere los criterios de selección de datos (consulte la Tarea 2.1) a la luz de la experiencia de modelado (es decir, la evaluación del modelo puede mostrar que se necesitan otros conjuntos de datos).
- Seleccione diferentes subconjuntos de datos (p. ej., diferentes atributos, solo datos que cumplan ciertas condiciones).
- Considere el uso de técnicas de muestreo (por ejemplo, una solución rápida puede implicar la reducción del tamaño del conjunto de datos de prueba o la herramienta puede no ser capaz de manejar el conjunto de datos completo, la prueba dividida y los conjuntos de datos de entrenamiento). También puede ser útil tener muestras ponderadas para dar diferente importancia a diferentes atributos o diferentes valores del mismo atributo.
- Documente la justificación de la inclusión/exclusión.
- Verifique las técnicas disponibles para el muestreo de datos.

¡Buena idea! Según los criterios de selección de datos, decida si uno o más atributos son más importantes que otros y pondere los atributos en consecuencia. Decida, en función del contexto (es decir, aplicación, herramienta, etc.), cómo manejar la ponderación.

3.2 Limpiar datos

Tarea

Limpiar datos

Elevar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de subconjuntos limpios de datos, la inserción de valores predeterminados adecuados o técnicas más ambiciosas, como la estimación de datos faltantes mediante modelado.

Producción

Informe de limpieza de datos

Describa las decisiones y acciones que se tomaron para abordar los problemas de calidad de los datos informados durante la tarea Verificar la calidad de los datos. El informe también debe abordar qué problemas de calidad de datos aún están pendientes si los datos se utilizarán en el ejercicio de extracción de datos y qué posibles efectos podrían tener en los resultados.

Actividades

- Reconsidere cómo lidiar con el tipo de ruido observado.
- Corrija, elimine o ignore el ruido.
- Decida cómo tratar los valores especiales y su significado. El área de valores especiales puede dar lugar a muchos resultados extraños y debe examinarse cuidadosamente. Podrían surgir ejemplos de valores especiales al tomar los resultados de una encuesta en la que no se hicieron ni respondieron algunas preguntas. Esto podría resultar en un valor de '99' para datos desconocidos. Por ejemplo, 99 para estado civil o afiliación política. También pueden surgir valores especiales cuando los datos se truncan, por ejemplo, '00' para personas de 100 años o todos los coches con 100.000 km en el reloj.
- Reconsidere los criterios de selección de datos (consulte la tarea 2.1) a la luz de las experiencias de limpieza de datos (es decir, es posible que desee incluir/excluir otros conjuntos de datos).

¡Buena idea!

Recuerde que algunos campos pueden ser irrelevantes para los objetivos de minería de datos y, por lo tanto, el ruido en esos campos no tiene importancia. Sin embargo, si se ignora el ruido por estos motivos, debe documentarse completamente, ya que las circunstancias pueden cambiar más adelante.

3.3 Construir datos

Tarea

Construir datos

Esta tarea incluye operaciones de preparación de datos constructivos como la producción de atributos derivados, nuevos registros completos o valores transformados para atributos existentes.

Actividades

- Consulte los mecanismos de construcción disponibles con la lista de herramientas sugeridas para el proyecto.
- Decida si es mejor realizar la construcción dentro o fuera de la herramienta (es decir, cuál es más eficiente, exacta, repetible).

- Reconsidere los criterios de selección de datos (consulte la Tarea 2.1) a la luz de las experiencias de construcción de datos (es decir, puede desear incluir/excluir otros conjuntos de datos).

Producción

Atributos derivados

Los atributos derivados son atributos nuevos que se construyen a partir de uno o más atributos existentes en el mismo registro. Un ejemplo podría ser área = largo * ancho.

¿Por qué debemos construir atributos derivados durante el curso de una investigación de minería de datos? No se debe pensar que solo los datos de bases de datos u otras fuentes son el único tipo de datos que se deben usar para construir un modelo. Los atributos derivados pueden construirse porque:

- El conocimiento previo nos convence de que algún hecho es importante y debe ser representado aunque actualmente no tengamos ningún atributo para representarlo.
- El algoritmo de modelado en uso maneja solo ciertos tipos de datos, por ejemplo, estamos usando regresión lineal y sospechamos que hay ciertas no linealidades que no se incluirán en el modelo.
- El resultado de la fase de modelado puede sugerir que ciertos hechos no están siendo cubiertos.

Actividades

Atributos derivados

- Decida si algún atributo debe normalizarse (p. ej., cuando se utiliza un algoritmo de agrupación con la edad y los ingresos en liras, predominará el ingreso).
- Considere agregar nueva información sobre la importancia relevante de los atributos agregando nuevos atributos (por ejemplo, ponderaciones de atributos, normalización ponderada).
- ¿Cómo se pueden construir o imputar los atributos que faltan? [Decidir el tipo de construcción (p. ej., agregado, promedio, inducción)].
- Agregue nuevos atributos a los datos accedidos.

¡Buena idea!

Antes de agregar atributos derivados, intente determinar si facilitan el proceso del modelo o facilitan el algoritmo de modelado y cómo lo hacen. Tal vez el "ingreso per cápita" sea un atributo mejor/más fácil de usar que el "ingreso per cápita". No derive atributos simplemente para reducir el número de atributos de entrada.

Otro tipo de atributo derivado son las transformaciones de atributo único, que normalmente se realizan para adaptarse a las necesidades de las herramientas de modelado.

Actividades **Transformaciones de un solo atributo**

- Especifique los pasos de transformación necesarios en términos de las instalaciones de transformación disponibles (por ejemplo, cambie la clasificación de un atributo numérico).
- Realizar pasos de transformación.

¡Pista! Las transformaciones pueden ser necesarias para transformar rangos en campos simbólicos (por ejemplo, edades en rangos de edad) o campos simbólicos ("definitivamente sí", "sí", "no sé", "no") en valores numéricos. Las herramientas de modelado o los algoritmos a menudo los requieren.

Producción **registros generados**

Los registros generados son registros completamente nuevos, que agregan nuevos conocimientos o representan nuevos datos que no están representados de otro modo, por ejemplo, habiendo segmentado los datos, puede ser útil generar un registro para representar el miembro prototípico de cada segmento para su posterior procesamiento.

Actividades Verifique las técnicas disponibles si es necesario (p. ej., mecanismos para construir prototipos para cada segmento de datos segmentados).

3.4 Integrar datos

Tarea **Integrar datos**

Estos son métodos mediante los cuales se combina la información de*múltiple* tablas u otras fuentes de información para crear nuevos registros o valores.

Producción **datos combinados**

Fusionar tablas se refiere a unir dos o más tablas que tienen información diferente sobre los mismos objetos. En esta etapa también puede ser recomendable generar nuevos registros. También se puede recomendar generar valores agregados.

La agregación se refiere a operaciones en las que se calculan nuevos valores al resumir información de múltiples registros y/o tablas.

Actividades

- Compruebe las instalaciones de integración si pueden integrar las fuentes de entrada según sea necesario.
- Integrar fuentes y almacenar resultados.
- Reconsidere los criterios de selección de datos (consulte la Tarea 2.1) a la luz de las experiencias de integración de datos (es decir, puede desear incluir/excluir otros conjuntos de datos).

¡Buena idea! Recuerde que algunos conocimientos pueden estar contenidos en formato no electrónico.

3.5 Formatear datos

Tarea

Formatear datos

Las transformaciones de formato se refieren principalmente a *sintáctico* modificaciones realizadas a los datos que no cambian su significado, pero que pueden ser requeridas por la herramienta de modelado.

Producción

Datos reformateados

Algunas herramientas tienen requisitos sobre el orden de los atributos, como que el primer campo sea un identificador único para cada registro o que el último campo sea el campo de resultado que el modelo debe predecir.

Actividades

Reorganización de atributos

Algunas herramientas tienen requisitos sobre el orden de los atributos, como que el primer campo sea un identificador único para cada registro o que el último campo sea el campo de resultado que el modelo debe predecir.

Reordenación de registros

Puede ser importante cambiar el orden de los registros en el conjunto de datos. Tal vez la herramienta de modelado requiera que los registros se clasifiquen según el valor del atributo de resultado.

Reformateado dentro del valor

- Estos son cambios puramente sintácticos realizados para satisfacer los requisitos de la herramienta de modelado específica.
- Reconsidere los criterios de selección de datos (consulte la Tarea 2.1) a la luz de las experiencias de limpieza de datos (es decir, puede desear incluir/excluir otros conjuntos de datos).

4 Modelado

4.1 Seleccionar técnica de modelado

Tarea **Seleccionar técnica de modelado**

Como primer paso en el modelado, seleccione la técnica de modelado real que se utilizará inicialmente. Si se aplican varias técnicas, realice esta tarea para cada técnica por separado.

No hay que olvidar que no todas las herramientas y técnicas son aplicables a todas y cada una de las tareas. Para ciertos problemas, solo algunas técnicas son apropiadas (consulte el Apéndice V.2 donde se analizan con más detalle las técnicas apropiadas para ciertos tipos de problemas de minería de datos). Entre estas herramientas y técnicas existen "Requisitos políticos" y otras restricciones, que limitan aún más las opciones disponibles para el minero. Puede ser que solo una herramienta o técnica esté disponible para resolver el problema en cuestión, e incluso entonces la herramienta puede no ser absolutamente la mejor desde el punto de vista técnico para el problema en cuestión.

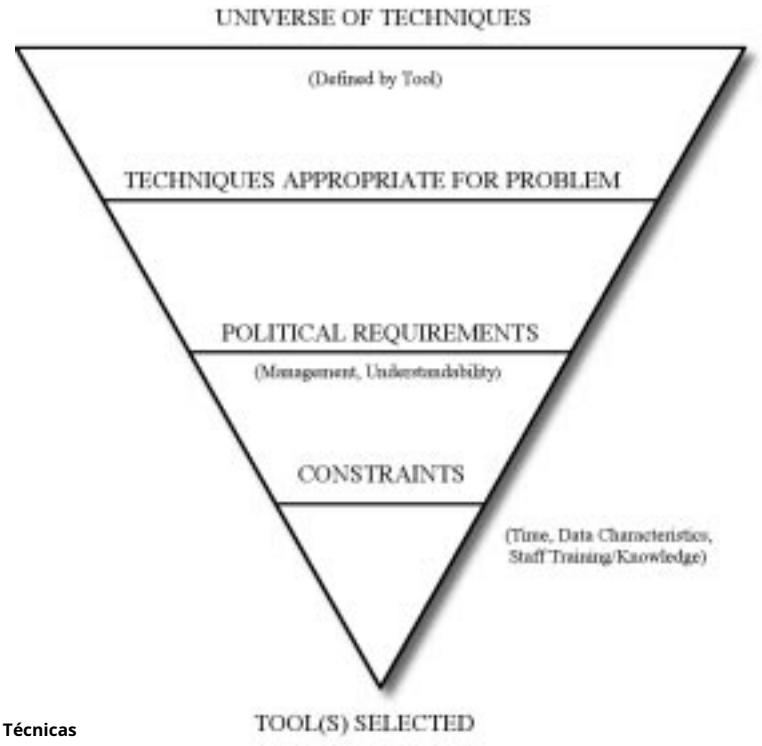


Figura 10:
Universo de Técnicas

Producción

Técnica de modelado

Registre la técnica de modelado real que se utiliza.

Actividades

Decidir la técnica adecuada para el ejercicio teniendo en cuenta la herramienta seleccionada.

Producción

Supuestos de modelado

Muchas técnicas de modelado hacen suposiciones específicas sobre los datos, la calidad de los datos o el formato de los datos.

Actividades

- Defina cualquier suposición incorporada hecha por la técnica sobre los datos (por ejemplo, calidad, formato, distribución).
- Compare estas suposiciones con las del Informe de descripción de datos.
- Asegúrese de que estas suposiciones se mantengan y retroceda a la Fase de preparación de datos si es necesario.

4.2 Generar diseño de prueba

Tarea

Generar diseño de prueba

Antes de construir un modelo, se debe definir un procedimiento para probar la calidad y validez del modelo. Por ejemplo, en tareas de minería de datos supervisadas como la clasificación, es común utilizar tasas de error como medidas de calidad para los modelos de minería de datos. Por lo tanto, el diseño de prueba especifica que el conjunto de datos debe separarse en conjunto de entrenamiento y de prueba, el modelo se construye sobre el conjunto de entrenamiento y su calidad se estima en el conjunto de prueba.

Producción

Diseño de prueba

Describa el plan previsto para entrenar, probar y evaluar los modelos. Un componente principal del plan es decidir cómo dividir el conjunto de datos disponible en datos de entrenamiento, datos de prueba y conjuntos de prueba de validación.

Actividades

- Verifique los diseños de prueba existentes para cada objetivo de minería de datos por separado.
- Decida los pasos necesarios (número de iteraciones, número de pliegues, etc.).
- Preparar los datos necesarios para la prueba.

4.3 Modelo de construcción

Tarea

Construir modelo

Ejecute la herramienta de modelado en el conjunto de datos preparado para crear uno o más modelos.

Producción

Configuración de parámetros

Con cualquier herramienta de modelado, a menudo hay una gran cantidad de parámetros que se pueden ajustar. Enumere los parámetros y sus valores elegidos, junto con la justificación de la elección.

Actividades

- Establecer parámetros iniciales.
- Documente las razones para elegir esos valores.

Producción

Modelos

Ejecute la herramienta de modelado en el conjunto de datos preparado para crear uno o más modelos.

Actividades

- Ejecute la técnica seleccionada en el conjunto de datos de entrada para producir el modelo.
- Resultados de la minería de datos de procesamiento posterior (p. ej., reglas de edición, árboles de visualización).

Producción

Descripcion del modelo

Describa el modelo resultante y evalúe su precisión esperada, robustez y posibles deficiencias. Informe sobre la interpretación de los modelos y las dificultades encontradas.

Actividades

- Describa cualquier característica del modelo actual que pueda ser útil para el futuro.
- Registre la configuración de los parámetros utilizados para producir el modelo.
- Proporcione una descripción detallada del modelo y cualquier característica especial.
- Para los modelos basados en reglas, enumere las reglas producidas más cualquier evaluación de la precisión y la cobertura del modelo general o por regla.
- Para los modelos opacos, enumere cualquier información técnica sobre el modelo (como la topología de la red neuronal) y las descripciones de comportamiento producidas por el proceso de modelado (como la precisión o la sensibilidad).
- Describa el comportamiento y la interpretación del modelo.
- Indique las conclusiones con respecto a los patrones en los datos (si los hay); a veces, el modelo revela hechos importantes sobre los datos sin un proceso de evaluación separado (p. ej., que el resultado o la conclusión se duplica en una de las entradas).

4.4 Modelo de evaluación

Tarea **Evaluar modelo**

Ahora se debe evaluar el modelo para garantizar que cumpla con los criterios de éxito de la minería de datos y que pase los criterios de prueba deseados. Esta es una evaluación puramente técnica basada en el resultado de las tareas de modelado.

Producción **Evaluación del modelo**

Resuma los resultados de esta tarea, enumere las cualidades de los modelos generados (por ejemplo, en términos de precisión) y clasifique su calidad en relación con los demás.

Actividades Evaluar el resultado con respecto a los criterios de evaluación

¡Buena idea! Se pueden construir "Tablas de elevación" y "Tablas de ganancia" para determinar qué tan bien está prediciendo el modelo.

- Resultado de la prueba de acuerdo con una estrategia de prueba (por ejemplo: Train and Test, Crossvalidation, bootstrapping, etc.).
- Comparar los resultados de la evaluación y la interpretación.
- Crear ranking de resultados con respecto al éxito y criterios de evaluación
- Seleccione los mejores modelos.
- Interpretar los resultados en términos comerciales (en la medida de lo posible en esta etapa).
- Obtenga comentarios sobre modelos por parte de expertos en datos o dominios.
- Comprobar la plausibilidad del modelo.
- Verifique los impactos para el objetivo de minería de datos.
- Verifique el modelo con la base de conocimientos dada para ver si la información descubierta es novedosa y útil.
- Comprobar la fiabilidad del resultado.
- Analizar los potenciales de despliegue de cada resultado.
- Si hay una descripción verbal del modelo generado (por ejemplo, a través de reglas), evalúe las reglas; ¿son lógicos, factibles, demasiados o pocos, ofenden el sentido común?
- Evaluar resultados.
- Obtenga información sobre por qué una determinada técnica de modelado y ciertas configuraciones de parámetros conducen a buenos o malos resultados.

Producción **Ajustes de parámetros revisados**

De acuerdo con la evaluación del modelo, revise la configuración de los parámetros y ajústelos para la próxima ejecución en la tarea 'Crear modelo'. Repita la construcción y evaluación de modelos hasta que encuentre el mejor modelo.

Actividades Ajuste los parámetros para dar un mejor modelo.

5 Evaluación

Los pasos de evaluación anteriores se ocuparon de factores como la precisión y la generalidad del modelo. Este paso evalúa el grado en que el modelo cumple con los objetivos comerciales y busca determinar si existe alguna razón comercial por la cual este modelo es deficiente. Compara los resultados con los criterios de evaluación definidos al inicio del proyecto.

Una buena manera de definir los resultados totales de un proyecto de minería de datos es usar la ecuación:

RESULTADOS = MODELOS + HALLAZGOS

En esta ecuación estamos definiendo que el resultado total del proyecto de minería de datos no son solo los modelos (aunque son, por supuesto, importantes) sino también los hallazgos que definimos como cualquier cosa*(aparte del modelo)*que es importante para cumplir los objetivos del negocio (o importante para generar nuevas preguntas, línea de enfoque o efectos secundarios (por ejemplo, problemas de calidad de datos descubiertos por el ejercicio de minería de datos). Nota: aunque el modelo está directamente relacionado con las preguntas del negocio, los hallazgos no necesitan estar relacionados con ninguna pregunta u objetivo, pero son importantes para el iniciador del proyecto.

5.1 Evaluar resultados

Tarea

Evaluar resultados

Los pasos de evaluación anteriores se ocuparon de factores como la precisión y la generalidad del modelo. Este paso evalúa el grado en que el modelo cumple con los objetivos comerciales y busca determinar si existe alguna razón comercial por la cual este modelo es deficiente. Otra opción de evaluación es probar los modelos en aplicaciones de prueba en la aplicación real si las limitaciones de tiempo y presupuesto lo permiten.

Además, la evaluación también evalúa otros resultados de minería de datos generados. Los resultados de la minería de datos cubren modelos que están necesariamente relacionados con los objetivos comerciales originales y todos los demás hallazgos que no están necesariamente relacionados con los objetivos comerciales originales, pero que también pueden revelar desafíos adicionales, información o sugerencias para direcciones futuras.

Producción

Evaluación de los resultados de la minería de datos con respecto a los criterios de éxito empresarial

Resuma los resultados de la evaluación en términos de criterios de éxito comercial, incluida una declaración final sobre si el proyecto ya cumple con los objetivos comerciales iniciales.

Actividades

- Comprender el resultado de la minería de datos.
- Interpretar los resultados en términos de la aplicación.
- Verifique los impactos para el objetivo de minería de datos.
- Verifique el resultado de la minería de datos con la base de conocimiento dada para ver si la información descubierta es novedosa y útil.
- Evaluar y evaluar el resultado con respecto a los criterios de éxito comercial. es decir, ¿ha logrado el proyecto los objetivos comerciales originales?
- Comparar los resultados de la evaluación y la interpretación.
- Crear ranking de resultados con respecto a criterios de éxito empresarial.
- Verifique los impactos del resultado para el objetivo de la aplicación inicial.
- ¿Existen nuevos objetivos comerciales que se abordarán más adelante en el proyecto o en nuevos proyectos?
- Establece conclusiones para futuros proyectos de minería de datos.

Producción

Modelos homologados

Después de la evaluación del modelo con respecto a los criterios de éxito comercial, eventualmente obtendrá modelos aprobados si los modelos generados cumplen con los criterios seleccionados.

5.2 Proceso de revisión

Tarea

Proceso de revisión

En este punto, el modelo resultante parece ser satisfactorio y parece satisfacer las necesidades comerciales. Ahora es apropiado hacer una revisión más exhaustiva del compromiso de minería de datos para determinar si hay algún factor o tarea importante que de alguna manera se haya pasado por alto. En esta etapa del ejercicio de minería de datos, la revisión del proceso adopta la forma de una revisión de control de calidad.

Producción

Revisión del proceso

Resuma la revisión del proceso y dé sugerencias para las actividades que se han perdido y/o deberían repetirse.

Actividades

- Dar una visión general del proceso de minería de datos utilizado
- Analizar el proceso de minería de datos

  Para cada etapa del proceso:

  - ¿Era necesario en retrospectiva?
  - ¿Se ejecutó de manera óptima?
- ¿De qué manera se podría mejorar?
- Identifica fallas.
- Identificar pasos engañosos.
- Identificar posibles acciones alternativas, caminos inesperados en el proceso.
- Revisar los resultados de la minería de datos con respecto a los criterios de éxito empresarial.

5.3 Determinar los próximos pasos

Tarea

Determinar los próximos pasos

De acuerdo con los resultados de la evaluación y la revisión del proceso, el proyecto decide cómo proceder en esta etapa. El proyecto debe decidir si finalizar este proyecto y pasar a la implementación o si iniciar más iteraciones o si configurar nuevos proyectos de minería de datos.

Producción

Lista de posibles acciones

Enumere posibles acciones adicionales junto con las razones a favor y en contra de cada opción.

Actividades

- Analizar el potencial de despliegue de cada resultado.
- Estimar el potencial de mejora del proceso actual.
- Compruebe los recursos restantes para determinar si permiten iteraciones de procesos adicionales (o si se pueden poner a disposición recursos adicionales).
- Recomendar continuaciones alternativas.
- Refinar el plan de proceso.

Producción

Decisión

Describa la decisión sobre cómo proceder junto con la justificación.

Actividades

- Clasifique las acciones posibles.
- Seleccione una de las acciones posibles.
- Documentar las razones de la elección.

6 Despliegue

6.1 Planificar el despliegue

Tarea **Planificar la implementación**

Esta tarea toma los resultados de la evaluación y concluye una estrategia para la implementación de los resultados de la minería de datos en el negocio.

Producción **Plan de empleo**

Resuma la estrategia de implementación, incluidos los pasos necesarios y cómo realizarlos.

Actividades

- Resuma los resultados desplegables.
- Desarrollar y evaluar planes alternativos para el despliegue.
- Decidir para cada resultado distinto de conocimiento o información.
- ¿Cómo se propagará el conocimiento o la información a sus usuarios?
- ¿Cómo se monitoreará el uso del resultado o se medirán sus beneficios (cuando corresponda)?
- Decida para cada modelo desplegable o resultado de software.
- ¿Cómo se implementará el resultado del modelo o software dentro de los sistemas de la organización?
- ¿Cómo se monitoreará su uso y se medirán sus beneficios (cuando corresponda)?
- Identificar posibles problemas al desplegar los resultados de la minería de datos (trampas del despliegue).

6.2 Seguimiento y mantenimiento del plan

Tarea **Monitoreo y mantenimiento del plan**

El monitoreo y el mantenimiento son temas importantes si el resultado de la minería de datos se convierte en parte del día a día del negocio y su entorno. Una preparación cuidadosa de una estrategia de mantenimiento ayuda a evitar períodos innecesariamente largos de uso incorrecto de los resultados de la minería de datos. Para monitorear el despliegue de los resultados de la minería de datos, el proyecto necesita un plan detallado sobre el proceso de monitoreo. Este plan tiene en cuenta el tipo específico de implementación.

Producción **Plan de seguimiento y mantenimiento**

Resumir la estrategia de monitoreo y mantenimiento, incluidos los pasos necesarios y cómo realizarlos.

Actividades

- Compruebe los aspectos dinámicos (es decir, ¿qué cosas podrían cambiar en el entorno?).
- ¿Cómo se controlará la precisión?
- ¿Cuándo no se debe volver a utilizar el resultado o modelo de minería de datos? ¿Identificar criterios (validez, umbral de precisión, nuevos datos, cambio en el dominio de la aplicación, etc.)? ¿Qué debería pasar si el modelo o el resultado ya no se pueden usar? (Actualizar modelo, configurar nuevo proyecto de minería de datos, etc.).
- ¿Los objetivos comerciales del uso del modelo cambiarán con el tiempo? Documente completamente el problema inicial que el modelo intentaba resolver.

- Desarrollar un plan de monitoreo y mantenimiento.

6.3 Producir informe final

Tarea

Producir informe final

Al final del proyecto, el líder del proyecto y su equipo redactan un informe final. Depende del plan de implementación, si este informe es solo un resumen del proyecto y sus experiencias o si este informe es una presentación final de los resultados de la minería de datos.

Producción

Reporte final

Al final del proyecto, habrá (al menos un) informe final donde se reúnan todos los hilos. Además de identificar los resultados obtenidos, el informe también debe describir el proceso, mostrar en qué costos se ha incurrido, definir cualquier desviación del plan original, describir los planes de implementación y hacer recomendaciones para el trabajo futuro. El contenido detallado real del informe depende en gran medida de la audiencia para el informe en particular.

Actividades

- Identifique qué informes se necesitan (presentación de diapositivas, resumen de gestión, hallazgos detallados, explicación de modelos, etc.).
- Analice qué tan bien se han cumplido los objetivos iniciales de minería de datos.
- Identificar grupos objetivo para el informe.
- Resuma la estructura y el contenido de los informes.
- Seleccione los hallazgos que se incluirán en los informes.
- Escribe un reporte.

Producción

Presentación final

Además de un Informe final, puede ser necesario hacer una Presentación final para resumir el proyecto, tal vez para el patrocinador de gestión, por ejemplo. La Presentación normalmente contiene un subconjunto de la información contenida en el Informe Final, pero estructurada de una manera diferente.

Actividades

- Decida el grupo objetivo para la presentación final (¿ya habrán recibido el informe final?).
- Seleccione qué elementos del informe final deben incluirse en la presentación final.

6.4 Proyecto de revisión

Tarea

Revisar proyecto

Evaluar lo que salió bien y lo que salió mal, lo que se hizo bien y lo que debe mejorarse.

Producción

Documentación de la experiencia

Resuma las experiencias importantes realizadas durante el proyecto. Por ejemplo, las trampas, los enfoques engañosos o las sugerencias para seleccionar las técnicas de minería de datos más adecuadas en situaciones similares podrían ser parte de esta documentación. En proyectos ideales, la documentación de la experiencia cubre también cualquier informe que haya sido escrito por miembros individuales del proyecto durante las fases del proyecto y sus tareas.

Actividades

- Entreviste a todas las personas significativas involucradas en el proyecto y pregúnteles sobre sus experiencias durante el proyecto.
- Si los usuarios finales de la empresa trabajan con los resultados de la minería de datos, entrevístelos: ¿están satisfechos? ¿Qué se podría haber hecho mejor? ¿Necesitan apoyo adicional?
- Resuma los comentarios y escriba la documentación de la experiencia.
- Analizar el proceso (cosas que funcionaron bien, errores cometidos, lecciones aprendidas, etc.).
- Documentar el proceso específico de minería de datos (¿Cómo se pueden retroalimentar los resultados y la experiencia de aplicación del modelo en el proceso?).
- Abstraerse de los detalles para que la experiencia sea útil para proyectos futuros.

IV Los resultados de CRISP-DM

Esta sección contiene breves descripciones del propósito y el contenido de los informes más importantes. Aquí, nos centramos en los informes destinados a comunicar los resultados de una fase a las personas que no participan en esta fase (y posiblemente no participan en este proyecto). Estas no son necesariamente idénticas a las salidas descritas en el modelo de referencia y la guía del usuario. El propósito de los productos es principalmente documentar los resultados.*mientras se realiza*el proyecto.

1 Comprensión empresarial

Los resultados de la fase de comprensión empresarial se pueden resumir en un informe. Sugerimos las siguientes secciones:

Fondo

Los antecedentes proporcionan una descripción general básica del contexto del proyecto. Esto enumera en qué área está trabajando el proyecto, qué problemas se han identificado y por qué la minería de datos parece proporcionar una solución.

Objetivos de negocio y criterios de éxito

Los objetivos comerciales describen cuáles son las metas del proyecto en términos comerciales. Para cada objetivo, se deben proporcionar los Criterios de Éxito Empresarial, es decir, medidas explícitas para determinar si el proyecto logró o no sus objetivos. Esta sección también debe enumerar los objetivos que se consideraron pero que se rechazaron. Se debe dar la justificación de la selección de objetivos.

Inventario de recursos

El Inventario de Recursos tiene como objetivo identificar el personal, las fuentes de datos, las instalaciones técnicas y otros recursos que puedan ser útiles para llevar a cabo el proyecto.

Requisitos, suposiciones y restricciones

Este resultado enumera los requisitos generales sobre cómo se ejecuta el proyecto, el tipo de resultados del proyecto, las suposiciones hechas sobre la naturaleza del problema y los datos que se utilizan y las restricciones impuestas al proyecto.

Riesgos y contingencias

Esta salida identifica los problemas que pueden ocurrir en el proyecto, describe las consecuencias y establece qué acción se puede tomar para minimizar el efecto.

Terminología

La Terminología permite que las personas que no están familiarizadas con los problemas que aborda el proyecto se familiaricen más con ellos.

Costos y beneficios

Esto describe los costos del proyecto y los beneficios comerciales previstos si el proyecto tiene éxito (por ejemplo, el retorno de la inversión). También deben destacarse otros beneficios menos tangibles (por ejemplo, la satisfacción del cliente).

Objetivos de minería de datos y criterios de éxito

Los objetivos de minería de datos establecen los resultados del proyecto que permiten el logro de los objetivos comerciales. Además de enumerar los enfoques de minería de datos probables, los criterios de éxito para los resultados también deben enumerarse en términos de minería de datos.

Plan de proyecto

Este enumera las etapas a ejecutar en el proyecto, junto con la duración, los recursos requeridos, las entradas, las salidas y las dependencias. Siempre que sea posible, debe hacer explícitas las iteraciones a gran escala en el proceso de extracción de datos, por ejemplo, repeticiones de las fases de modelado y evaluación.

Evaluación inicial de herramientas y técnicas.

Esta sección ofrece una visión inicial de qué herramientas y técnicas es probable que se utilicen y cómo. Describe los requisitos para las herramientas y técnicas, enumera las herramientas y técnicas disponibles y las relaciona con los requisitos.

2 Comprensión de datos

Los resultados de la fase de comprensión de datos suelen documentarse en varios informes. Idealmente, estos informes deben ser escritos mientras se realizan las tareas respectivas. Los informes describen los conjuntos de datos que se exploran durante la comprensión de los datos. Para el informe final es suficiente un resumen de las partes más relevantes.

Informe inicial de recopilación de datos

Este informe describe cómo se capturaron y extrajeron las diferentes fuentes de datos identificadas en el inventario.

Temas a tratar:

- Fondo de datos.
  - Lista de fuentes de datos con una amplia área de datos requeridos cubiertos por cada uno.
  - Para cada fuente de datos, método de adquisición o extracción.
  - Problemas encontrados en la adquisición o extracción de datos.

Informe de descripción de datos

Se describe cada conjunto de datos adquirido.

Temas a tratar:

- Cada fuente de datos se describe en detalle.
  - Lista de tablas (puede ser solo una) u otros objetos de la base de datos.
  - Descripción de cada campo incluyendo unidades, códigos utilizados, etc.

Informe de exploración de datos

Describe la exploración de datos y sus resultados.

Temas a tratar:

- Antecedentes que incluyen amplios objetivos de exploración de datos. Para cada área de exploración realizada:
  - Regularidades o patrones esperados.
  - Método de detección.
  - Regularidades o patrones encontrados, esperados e inesperados.
  - Cualquier otra sorpresa.
  - Conclusiones para la transformación de datos, limpieza de datos y cualquier otro preprocesamiento.
  - Conclusiones relacionadas con las metas de minería de datos u objetivos comerciales.
  - Resumen de conclusiones.

Informe de calidad de datos

Este informe describe la integridad y precisión de los datos.

Temas a tratar:

- Antecedentes que incluyen amplias expectativas sobre la calidad de los datos. Para cada conjunto de datos:
  - Enfoque adoptado para evaluar la calidad de los datos.
  - Resultados de la evaluación de la calidad de los datos.
  - Resumen de las conclusiones sobre la calidad de los datos.

3 Preparación de datos

Los informes en la fase de preparación de datos se centran en los pasos de preprocesamiento que producen los datos que se extraerán.

Informe de descripción del conjunto de datos

Esto proporciona una descripción del conjunto de datos (después del procesamiento previo) y el proceso mediante el cual se produjo.

Temas a tratar:

- Antecedentes que incluyen objetivos generales y un plan para el procesamiento previo.
- Justificación para la inclusión/exclusión de conjuntos de datos. Para cada conjunto de datos incluido:
  - Descripción del procesamiento previo, incluidas las acciones que fueron necesarias para abordar cualquier problema de calidad de los datos.
  - Descripción detallada del conjunto de datos resultante, tabla por tabla y campo por campo.
  - Justificación para la inclusión/exclusión de atributos.
  - Descubrimientos realizados durante el preprocesamiento y cualquier implicación para trabajos posteriores.
  - Resumen y conclusiones.

4 Modelado

Los resultados producidos durante la fase de modelado se pueden combinar en un solo informe. Sugerimos las siguientes secciones:

Suposición de modelado

Esta sección define *explícitamente* cualquier suposición hecha sobre los datos y cualquier suposición que esté implícita en la técnica de modelado que se utilizará.

Diseño de prueba

Esta sección describe cómo se construyen, prueban y evalúan los modelos.

Temas a tratar:

- Antecedentes: describe el modelado realizado y su relación con los objetivos de la minería de datos.

  Para cada tarea de modelado:
  - Descripción amplia del tipo de modelo y los datos de entrenamiento a utilizar.
  - Explicación de cómo se probará o evaluará el modelo.
  - Descripción de cualquier dato requerido para la prueba.
  - Plan para la producción de datos de prueba, si los hubiere.
  - Descripción de cualquier examen planificado de modelos por parte de expertos en dominios o datos.
  - Resumen del plan de prueba.

Descripcion del modelo

Este informe describe los modelos entregados y resume el proceso por el cual fueron producidos.

Temas a tratar:

- Descripción general de los modelos producidos. Para cada modelo:
  - Tipo de modelo y relación con los objetivos de minería de datos.
  - Ajustes de parámetros utilizados para producir el modelo.
  - Descripción detallada del modelo y sus características especiales.
    Por ejemplo:
    - Para los modelos basados en reglas, enumere las reglas producidas más cualquier evaluación de la precisión y la cobertura del modelo general o por regla.
    - Para los modelos opacos, enumere cualquier información técnica sobre el modelo (como la topología de la red neuronal) y las descripciones de comportamiento producidas por el proceso de modelado (como la precisión o la sensibilidad).
    - Descripción del comportamiento del modelo e interpretación.
  - Conclusiones con respecto a los patrones en los datos (si los hay); a veces, el modelo revelará hechos importantes sobre los datos sin un proceso de evaluación separado (p. ej., que el resultado o la conclusión se duplica en una de las entradas).
- Resumen de conclusiones.

Evaluación del modelo

Esta sección describe los resultados de probar los modelos de acuerdo con el diseño de prueba.

Temas a tratar:

- Descripción general del proceso de evaluación y los resultados, incluidas las desviaciones del plan de prueba.
  Para cada modelo:
  - Evaluación detallada del modelo, incluidas medidas como la precisión y la interpretación del comportamiento.
  - Cualquier comentario sobre modelos por parte de expertos en dominios o datos.
  - Evaluación resumida del modelo.
  - Información sobre por qué una determinada técnica de modelado y ciertas configuraciones de parámetros dieron buenos o malos resultados.
  - Evaluación resumida del conjunto completo de modelos.

5 Evaluación

Evaluación de los resultados de la minería de datos con respecto a los criterios de éxito empresarial Este informe compara los resultados de la minería de datos con los objetivos comerciales y los criterios de éxito empresarial.

Temas a tratar:

- Revisión de los objetivos comerciales y los criterios de éxito comercial (que pueden haber cambiado durante y/o como resultado de la extracción de datos).
  Para cada Criterio de Éxito Empresarial:
  - Comparación detallada entre el criterio de éxito y los resultados de la minería de datos.
  - Conclusiones sobre la posibilidad de alcanzar el criterio de éxito y la idoneidad del proceso de minería de datos.
- revisión del éxito del proyecto; ¿Ha logrado el proyecto los objetivos comerciales originales?
- ¿Existen nuevos objetivos comerciales que se abordarán más adelante en el proyecto o en nuevos proyectos?
- Conclusiones para futuros proyectos de minería de datos.

Revisión del proceso

Esta sección evalúa la eficacia del proyecto e identifica los factores que pueden haberse pasado por alto y que deben tenerse en cuenta si se repite el proyecto.

Lista de posibles acciones

Esta sección hace recomendaciones con respecto a los próximos pasos en el proyecto.

6 Despliegue

Plan de empleo

Esta sección especifica la implementación de los resultados de la minería de datos.

Temas a tratar:

- Resumen de los resultados implementables (derivados del informe Próximos pasos).
- Descripción del plan de despliegue.

Plan de seguimiento y mantenimiento

El plan de seguimiento y mantenimiento especifica cómo se mantendrán los resultados desplegados.

Temas a tratar:

- Descripción general del despliegue de resultados e indicación de qué resultados pueden requerir actualización (y por qué).

  Para cada resultado desplegado:

  - Descripción de cómo se activará la actualización (actualizaciones periódicas, evento desencadenante, supervisión del rendimiento).
  - Descripción de cómo se realizará la actualización.
- Resumen del proceso de actualización de resultados.

Reporte final

El informe final se utiliza para resumir el proyecto y sus resultados.

Contenido:

- Resumen de Entendimiento Empresarial: antecedentes, objetivos y criterios de éxito.
- Resumen del proceso de minería de datos.
- Resumen de los resultados de la minería de datos.
- Resumen de evaluación de resultados.
- Resumen de los planes de despliegue y mantenimiento.
- Análisis coste-beneficio.
- Conclusiones para el negocio.
- Conclusiones para la futura minería de datos.

7 Resumen de dependencias

La siguiente tabla resume los principales insumos para los entregables. Esto no significa que solo se deban considerar las entradas enumeradas; por ejemplo, los objetivos comerciales deben ser generalizados para todos los entregables. Sin embargo, los entregables deben abordar cuestiones específicas planteadas por sus aportes.

Phase	Deliverable	Refers To	Closely Related To
Business Understanding	Background		
	Business Objectives	Background	Terminology
	Business Success Criteria	Business Objectives	
	Inventory of Resources		
	Requirements, Assumptions & Constraints	Business Objectives	
	Risks & Contingencies	Business Objectives; Business Success Criteria	
	Terminology	Background	Business Objectives
	Costs & Benefits	Business Objectives	Project Plan
	Data Mining Goals	Business Objectives; Requirements, Assumptions & Constraints	
	Data Mining Success Criteria	Business Success Criteria; Requirements, Assumptions & Constraints; Data Mining Goals	
	Project Plan	Business Objectives; Inventory of Resources; Requirements, Assumptions & Constraints; Risks & Contingencies	Costs & Benefits
Data Understanding	Initial Data Collection Report	Business Goals; Inventory of Resources; Data Mining Goals	
	Data Description Report	Business Goals; Initial Data Collection Report	Data Quality Report
	Data Quality Report	Business Goals; Initial Data Collection Report	Data Description Report
	Exploratory Analysis Report	Business Goals; Data Description Report; Data Quality Report	
Data Preparation	Data Set & Data Set Description	Business Goals; Data Mining Goals; Data Description Report; Data Quality Report; Exploratory Analysis Report	
Modeling	Test Design	Data Mining Goals; Data Mining Success Criteria	
	Models	Data Mining Goals	Parameter Settings
	Parameter Settings	Data Mining Goals	Models
	Model Description	Models; Parameter Settings, Test Design	
	Assessment	Data Mining Success Criteria; Test Design; Models	
Evaluation	Assessment w.r.t. Business Success Criteria	Business Success Criteria; Terminology	
	Review of Process	Business Goals; Assessment w.r.t. Business Success Criteria	
	Next Steps	Project Plan; Assessment w.r.t. Business Success Criteria	
Deployment	Deployment Plan	Business Goals; Requirements, Assumptions & Constraints	Maintenance Plan
	Maintenance Plan	Business Goals; Requirements, Assumptions & Constraints	Deployment Plan
	Final Report & Presentation	Business Goals; Terminology; Assessment w.r.t. Business Success Criteria	
	Experience Documentation	Project Plan; Review of Process	

V Apéndice

1 Glosario/terminología

Actividad

Parte de una tarea en la Guía del usuario, describe acciones para realizar una tarea.

Metodología CRISP-DM

El término general para todos los conceptos desarrollados y definidos en CRISP-DM.

Contexto de minería de datos

Conjunto de restricciones y suposiciones tales como tipo de problema, técnicas o herramientas, dominio de aplicación.

Tipo de problema de minería de datos

Clase de problemas típicos de minería de datos, como descripción y resumen de datos, segmentación, descripciones de conceptos, clasificación, predicción, análisis de dependencia.

Genérico

Una tarea que se mantiene en todos los proyectos de minería de datos posibles, tan completa, es decir, cubre todo el proceso de minería de datos y todas las aplicaciones de minería de datos posibles y estable, es decir, válida para desarrollos aún no previstos como nuevas técnicas de modelado, como sea posible.

Modelo

Capacidad de aplicar a un conjunto de datos para predecir un atributo objetivo, ejecutable.

Producción

Resultado tangible de realizar una tarea.

Fase

Término de alto nivel para parte del modelo de proceso, consta de tareas relacionadas.

Instancia de proceso

Un proyecto específico descrito en términos del modelo de proceso.

Modelo de proceso

Define la estructura de los proyectos de minería de datos y brinda orientación para su ejecución, consta de modelo de referencia y guía de usuario.

Modelo de referencia

Descomposición de proyectos de minería de datos en fases, tareas y salidas.

Especializado

Una tarea que hace suposiciones específicas en contextos específicos de minería de datos.

Tarea

Parte de una fase, serie de actividades para producir uno o más productos.

Guía del usuario

Asesoramiento específico sobre cómo realizar proyectos de minería de datos.

2 tipos de problemas de minería de datos

Por lo general, el proyecto de minería de datos implica una combinación de diferentes tipos de problemas, que juntos resuelven el problema empresarial.

2.1 Descripción y resumen de datos

*Descripción y resumen de datos*tiene como objetivo la descripción concisa de las características de los datos, típicamente en forma elemental y agregada. Esto le da al usuario una visión general de la estructura de los datos. A veces, la descripción y el resumen de datos por sí solos pueden ser un objetivo de un proyecto de minería de datos. Por ejemplo, un minorista podría estar interesado en la facturación de todos los puntos de venta desglosados por categorías. Los cambios y las diferencias con respecto a un período anterior se pueden resumir y resaltar. Este tipo de problema estaría en el extremo inferior de la escala de problemas de minería de datos.

Sin embargo, en casi todos los proyectos de minería de datos, la descripción y el resumen de datos es un objetivo secundario en el proceso, generalmente en las primeras etapas. Al comienzo de un proceso de minería de datos, el usuario a menudo no conoce ni el objetivo preciso del análisis ni la naturaleza precisa de los datos. El análisis exploratorio inicial de datos puede ayudar a comprender la naturaleza de los datos y encontrar posibles hipótesis para la información oculta. Las técnicas de visualización y estadísticas descriptivas simples proporcionan una primera percepción de los datos. Por ejemplo, la distribución de la edad de los clientes y sus áreas de vivienda da pistas sobre qué partes de un grupo de clientes deben abordarse mediante estrategias de marketing adicionales.

La descripción y el resumen de datos normalmente se producen en combinación con otros tipos de problemas de minería de datos. Por ejemplo, la descripción de datos puede conducir a la postulación de segmentos interesantes en los datos. Una vez que se identifican y definen los segmentos, es útil una descripción y un resumen de estos segmentos. Es recomendable realizar la descripción y el resumen de los datos antes de abordar cualquier otro tipo de problema de minería de datos. En este documento, esto se refleja en el hecho de que la descripción y resumen de datos es una tarea en la fase de comprensión de datos.

El resumen también juega un papel importante en la presentación de los resultados finales. Los resultados de los otros tipos de problemas de minería de datos (por ejemplo, descripciones de conceptos o modelos de predicción) también pueden considerarse resúmenes de datos, pero en un nivel conceptual más alto.

Muchos sistemas de informes, paquetes estadísticos, sistemas OLAP y EIS pueden cubrir la descripción y el resumen de datos, pero por lo general no proporcionan ningún método para realizar un modelado más avanzado. Si la descripción y el resumen de datos se considera un tipo de problema independiente y no se requiere más modelado, estas herramientas también son adecuadas para llevar a cabo compromisos de minería de datos.

2.2 Segmentación

El tipo de problema de minería de datos*segmentación*apunta a la separación de los datos en subgrupos o clases interesantes y significativos. Todos los miembros de un subgrupo comparten características comunes. Por ejemplo, en el análisis de la cesta de la compra se podrían definir segmentos de cestas en función de los artículos que contengan.

La segmentación se puede realizar de forma manual o (semi)automática. El analista puede formular hipótesis sobre ciertos subgrupos como relevantes para la pregunta comercial en función del conocimiento previo o en función del resultado de la descripción y el resumen de datos. Sin embargo, también existen técnicas de agrupamiento automático que pueden detectar estructuras previamente insospechadas y ocultas en los datos que permiten la segmentación.

La segmentación puede ser un tipo de problema de minería de datos en sí mismo. Entonces la detección de segmentos sería el objetivo principal de la minería de datos. Por ejemplo, todas las direcciones en áreas de código postal con edad e ingresos superiores al promedio pueden seleccionarse para enviar anuncios por correo sobre seguros de enfermería domiciliaria.

A menudo, sin embargo, muy a menudo la segmentación es un paso hacia la solución de otros tipos de problemas. Entonces, el propósito puede ser mantener manejable el tamaño de los datos o encontrar subconjuntos de datos homogéneos que sean más fáciles de analizar. Por lo general, en grandes conjuntos de datos, varias influencias se superponen y oscurecen los patrones interesantes. Luego, la segmentación adecuada facilita la tarea. Por ejemplo, analizar las dependencias entre artículos en millones de cestas de la compra es muy difícil. Es mucho más fácil (y más significativo, por lo general) identificar dependencias en segmentos interesantes de cestas de la compra, por ejemplo, cestas de alto valor, cestas que contienen artículos de conveniencia o cestas de un día u hora en particular.

Nota: En la literatura existe una confusión de términos. La segmentación a veces se denomina agrupación o clasificación. El último término es confuso porque algunas personas lo usan para referirse a la creación de clases, mientras que otros se refieren a la creación de modelos para predecir clases conocidas para casos nunca antes vistos. En este documento, restringimos el término clasificación al último significado (ver más abajo) y usamos el término segmentación para el primer significado, aunque se pueden usar técnicas de clasificación para obtener descripciones de los segmentos descubiertos.

Técnicas apropiadas:

- Técnicas de agrupamiento.
- Redes neuronales.
- Visualización.

Ejemplo:

Una empresa de automóviles recopila periódicamente información sobre sus clientes en relación con sus características socioeconómicas, como ingresos, edad, sexo, profesión, etc. Mediante el análisis de conglomerados, la empresa puede dividir a sus clientes en subgrupos más comprensibles y analizar la estructura de cada subgrupo. Las estrategias de marketing específicas se implementan para cada grupo por separado.

2.3 Descripciones de conceptos

*Descripción del concepto*tiene como objetivo un*comprensible*descripción de conceptos o clases. El propósito no es desarrollar modelos completos con alta precisión de predicción, sino obtener información. Por ejemplo, una empresa puede estar interesada en saber más sobre sus clientes leales y desleales. A partir de una descripción conceptual de estos conceptos (clientes leales y desleales), la empresa podría inferir qué se podría hacer para mantener a los clientes leales o para transformar a los clientes desleales en clientes leales.

La descripción del concepto tiene una estrecha conexión tanto con la segmentación como con la clasificación. La segmentación puede dar lugar a una enumeración de objetos pertenecientes a un concepto o clase sin ninguna descripción comprensible. Por lo general, hay una segmentación antes de realizar la descripción del concepto. Algunas técnicas, por ejemplo, las técnicas de agrupamiento conceptual, realizan la segmentación y la descripción de conceptos al mismo tiempo.

Las descripciones de conceptos también se pueden utilizar con fines de clasificación. Por otro lado, algunas técnicas de clasificación producen modelos de clasificación comprensibles, que luego pueden considerarse como descripciones de conceptos. La distinción importante es que la clasificación pretende ser completa en algún sentido. El modelo de clasificación debe aplicarse a*todo*casos en la población seleccionada. Por otro lado, las descripciones de los conceptos no necesitan ser completas. Es suficiente si describen partes importantes de los conceptos o clases. En el ejemplo anterior, puede ser suficiente obtener descripciones conceptuales de aquellos clientes que son claramente leales.

Técnicas apropiadas:

- Métodos de inducción de reglas.
- Agrupamiento conceptual.

Ejemplo:

Usando datos sobre los compradores de autos nuevos y usando una técnica de inducción de reglas, una compañía de automóviles podría generar reglas que describan a sus clientes leales y desleales. A continuación se muestran ejemplos de las reglas generadas:

Si SEXO = masculino y EDAD > 51 SEXO = luego CLIENTE = leal
Si femenino y EDAD > 21 PROFESIÓN = gerente luego CLIENTE = leal
Si y EDAD < 51 ESTADO FAMILIAR = soltero y luego CLIENTE = desleal
Si EDAD < 51 luego CLIENTE = desleal

2.4 Clasificación

*Clasificación*asume que hay un conjunto de objetos, caracterizados por algunos atributos o características, que pertenecen a diferentes clases. La etiqueta de clase es un valor discreto (simbólico) y se conoce para cada objeto. El objetivo es construir modelos de clasificación (a veces llamados clasificadores), que asignan la etiqueta de clase correcta a objetos no vistos ni etiquetados previamente. Los modelos de clasificación se utilizan principalmente para el modelado predictivo.

Las etiquetas de clase se pueden dar por adelantado, por ejemplo definidas por el usuario o derivadas de la segmentación.

La clasificación es uno de los tipos de problemas de minería de datos más importantes que ocurre en una amplia gama de aplicaciones. Muchos problemas de minería de datos pueden transformarse en problemas de clasificación. Por ejemplo, la calificación crediticia trata de evaluar el riesgo crediticio de un nuevo cliente. Esto se puede transformar en un problema de clasificación creando dos clases, buenos y malos clientes. Se puede generar un modelo de clasificación a partir de los datos de clientes existentes y su comportamiento crediticio. Este modelo de clasificación se puede usar para asignar un nuevo cliente potencial a una de las dos clases y, por lo tanto, aceptarlo o rechazarlo.

La clasificación tiene conexiones con casi todos los demás tipos de problemas. Los problemas de predicción pueden transformarse en problemas de clasificación agrupando etiquetas de clases continuas, ya que las técnicas de agrupamiento permiten transformar rangos continuos en intervalos discretos. Estos intervalos discretos se usan luego como etiquetas de clase en lugar de los valores numéricos exactos y, por lo tanto, conducen a un problema de clasificación. Algunas técnicas de clasificación producen descripciones comprensibles de clases o conceptos. También existe una conexión con el análisis de dependencia porque los modelos de clasificación suelen explotar y dilucidar las dependencias entre atributos.

La segmentación puede proporcionar las etiquetas de clase o restringir el conjunto de datos de modo que se puedan construir buenos modelos de clasificación.

Es útil analizar las desviaciones antes de construir un modelo de clasificación. Las desviaciones y los valores atípicos pueden oscurecer los patrones que permitirían un buen modelo de clasificación. Por otro lado, un modelo de clasificación también se puede utilizar para identificar desviaciones y otros problemas con los datos.

Técnicas apropiadas:

- Análisis discriminante.
- Métodos de inducción de reglas.
- Aprendizaje de árboles de decisión.
- Redes neuronales.
- K Vecino más cercano.
- Razonamiento basado en casos.
- Algoritmos genéticos.

Ejemplo:

Los bancos generalmente cuentan con información sobre el comportamiento de pago de sus solicitantes de crédito. Combinando esta información financiera con otra información sobre los clientes como sexo, edad, ingresos, etc., es posible desarrollar un sistema para clasificar a los nuevos clientes como buenos o malos clientes (es decir, el riesgo de crédito en la aceptación de un cliente es bajo o alto, respectivamente).

2.5 Predicción

Otro tipo de problema importante que ocurre en una amplia gama de aplicaciones es*predicción*. La predicción es muy similar a la clasificación. La única diferencia es que en la predicción el atributo objetivo (clase) no es un atributo cualitativo discreto sino continuo. El objetivo de la predicción es encontrar el valor numérico del atributo objetivo para objetos invisibles. En la literatura, este tipo de problema a veces se denomina regresión. Si la predicción trata con datos de series de tiempo, a menudo se le llama pronóstico.

Técnicas apropiadas:

- Análisis de regresión.
- Árboles de regresión.
- Redes neuronales.
- K Vecino más cercano.
- Métodos de Box-Jenkins.
- Algoritmos genéticos.

Ejemplo:

Los ingresos anuales de una empresa internacional se correlacionan con otros atributos como la publicidad, el tipo de cambio, la tasa de inflación, etc. Con estos valores (o sus estimaciones fiables para el próximo año), la empresa puede predecir sus ingresos esperados para el próximo año.

2.6 Análisis de dependencia

*Análisis de dependencia*consiste en encontrar un modelo que describa dependencias significativas (o asociaciones) entre elementos de datos o eventos. Las dependencias se pueden usar para predecir el valor de un elemento de datos dada la información sobre otros elementos de datos. Aunque las dependencias se pueden usar para el modelado predictivo, se usan principalmente para la comprensión. Las dependencias pueden ser estrictas o probabilísticas.

Las asociaciones son un caso especial de dependencias, que recientemente se han vuelto muy populares. Las asociaciones describen afinidades de elementos de datos (es decir, elementos de datos o eventos que frecuentemente ocurren juntos). Un escenario de aplicación típico para asociaciones es el análisis de cestas de la compra. Allí, una regla como “en el 30 por ciento de todas las compras, la cerveza y el maní se compraron juntos” es un ejemplo típico de una asociación.

Los algoritmos para detectar asociaciones son muy rápidos y producen muchas asociaciones. Seleccionar los más interesantes es un desafío.

El análisis de dependencia tiene conexiones cercanas con la predicción y la clasificación, donde las dependencias se usan implícitamente para la formulación de modelos predictivos. También hay una conexión con las descripciones de conceptos, que a menudo resaltan las dependencias.

En las aplicaciones, el análisis de dependencias suele coincidir con la segmentación. En grandes conjuntos de datos, las dependencias rara vez son significativas porque muchas influencias se superponen entre sí. En tales casos, es recomendable realizar un análisis de dependencia en segmentos más homogéneos de los datos.

*patrones secuenciales*son un tipo especial de dependencias donde se considera el orden de los eventos. En el dominio de la cesta de la compra, las asociaciones describen dependencias entre artículos en un momento dado. Los patrones secuenciales describen los patrones de compra de un cliente en particular o de un grupo de clientes a lo largo del tiempo.

Técnicas Apropiadas:

- Análisis de correlación.
- Análisis de regresión.
- Reglas de asociación.
- Redes bayesianas.
- Programación Lógica Inductiva.
- Técnicas de visualización.

Ejemplo 1:

Usando el análisis de regresión, un analista de negocios descubrió que existe una dependencia significativa entre las ventas totales de un producto y su precio y el monto de los gastos totales para el anuncio. Una vez que el analista descubre este conocimiento, puede alcanzar el nivel deseado de ventas cambiando el precio y/o el gasto publicitario en consecuencia.

Ejemplo 2:

Al aplicar algoritmos de reglas de asociación a datos sobre accesorios de automóviles, una empresa de automóviles descubrió que si se pide una radio, también se pide una caja de cambios automática en el 95 por ciento de los casos. En base a esta dependencia, la compañía automotriz decide ofrecer estos accesorios como una combinación que conduce a la reducción de costos.

La minería de datos marca la diferencia

SPSS Inc. permite a las organizaciones desarrollar relaciones más rentables con los clientes al proporcionar soluciones analíticas que descubren lo que quieren los clientes y predicen lo que harán. La empresa ofrece soluciones analíticas en la intersección de la gestión de relaciones con los clientes y la inteligencia comercial. Las soluciones analíticas de SPSS integran y analizan datos de mercado, de clientes y operativos y brindan resultados en mercados verticales clave en todo el mundo, incluidos: telecomunicaciones, atención médica, banca, finanzas, seguros, manufactura, venta minorista, bienes de consumo empaquetados, investigación de mercado y el sector público. Para más información visite **www.spss.com.**

Cómo ponerse en contacto con SPSS

Para obtener más información, llame a la oficina de SPSS más cercana o visite nuestro sitio web en la World Wide Web en **www.spss.com**

SPSS Inc.	+ 1.312.651.3000
gratuito	+ 1.800.543.2185
SPSS Argentina	+ 5411.4814.5030
SPSS Asia Pacífico	+ 65.245.9110
SPSS Australia	+ 61.2.9954.5660
gratuito	+ 1.800.024.836
SPSS Bélgica	+ 32.16.317070
SPSS Benelux	+ 31.183.651.777
SPSS Brasil	+ 55.11.5505.3644
SPSS República Checa	+ 420.2.24813839
SPSS Dinamarca	+ 45.45.46.02.00
SPSS África Oriental	+ 254.2.577.262
Sistemas federales SPSS (EE. UU.)	+ 1.703.527.6777
gratuito	+ 1.800.860.5762
SPSS Finlandia	+ 358.9.4355.920
SPSS Francia	+ 01.55.35.27.00
SPSS Alemania	+ 49.89.4890740
SPSS Hellas	+ 30.1.72.51.925
SPSS Hispanoportuguesa	+ 34.91.447.37.00
SPSS Hong Kong	+ 852.2.811.9662
SPSS Irlanda	+ 353.1.415.0234
SPSS España	+ 972.9.9526700
SPSS Italia	+ 39.051.252573
SPSS Japón	+ 81.3.5466.5511
SPSS Corea	+ 82.2.3446.7651
SPSS América Latina	+ 1.312.651.3539
SPSS Malasia	+ 603.7873.6477
SPSS México	+ 52.5.682.87.68
SPSS Miami	+ 1.305.627.5700
SPSS Noruega	+ 47.22.40.20.60
SPSS Polonia	+ 48.12.6369680
SPSS Rusia	+ 7.095.125.0069
SPSS Suiza	+ 41.1.266.90.30
SPSS Singapur	+ 65.324.5150
SPSS Sudáfrica	+ 27.11.807.3189
SPSS Asia del Sur	+ 91.80.2088069
SPSS Suecia	+ 46.8.506.105.50
SPSS Taiwán	+ 886.2.25771100
SPSS Tailandia	+ 66.2.260.7070
SPSS Reino Unido	+ 44.1483.719200

SPSS es una marca comercial registrada y los demás productos de SPSS mencionados son marcas comerciales de SPSS Inc. Todos los demás nombres son marcas comerciales de sus respectivos propietarios.

Impreso en EE. UU. © Copyright 2000 SPSS Inc. CRISPWP-0800

SPSS®