



BIG DATA & DATA SCIENCE

PROGRAMA DE ESPECIALIZACIÓN



Business Data Discovery

Empezar con el negocio





Temas



Procesos de Comprensión
Analítica



Entendimiento de Negocio



Recolección de Datos



Generación de Análisis



Medición de KPIs

Procesos de Comprensión Analítica

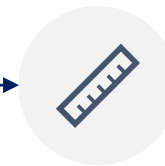


DATOS

INSIGHT

ACCIÓN

IMPACTO



ADQUIRIR

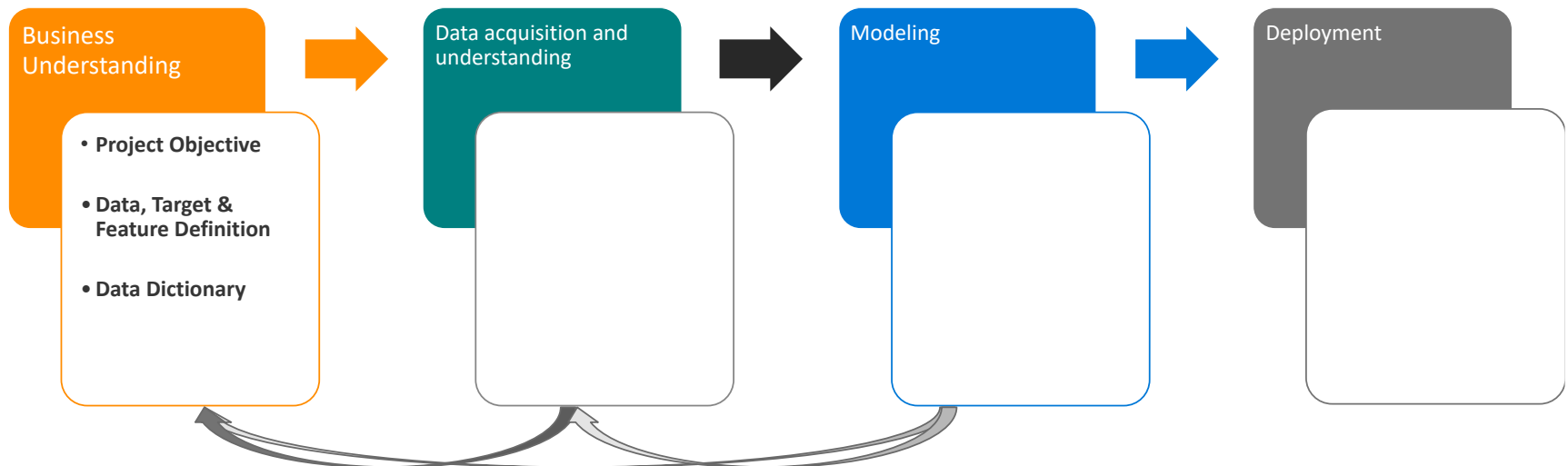
ORGANIZAR

ANALIZAR

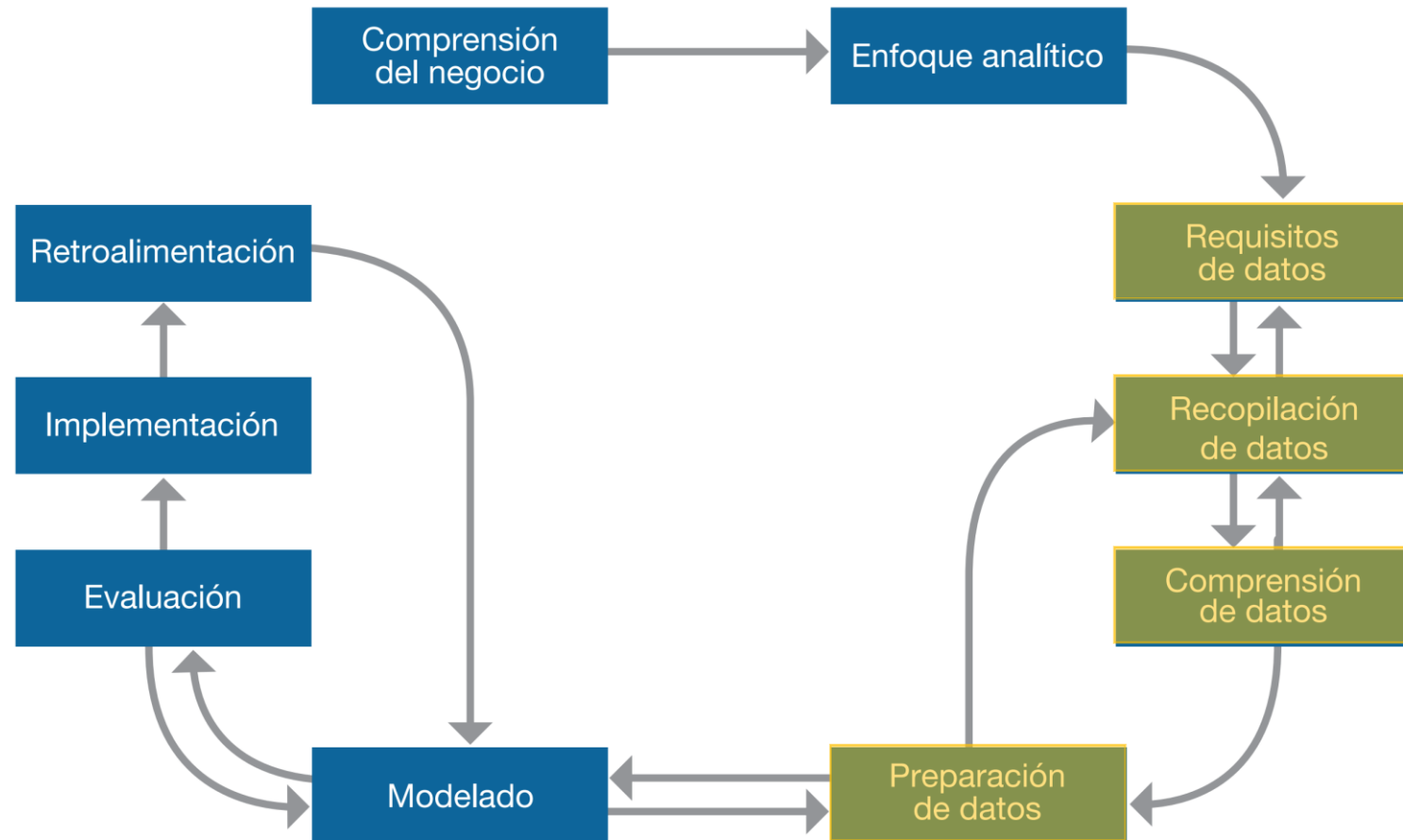
ENTREGAR

MEDIR

Las etapas del ciclo de vida de TDSP pueden integrarse con entregables y puntos de control específicos

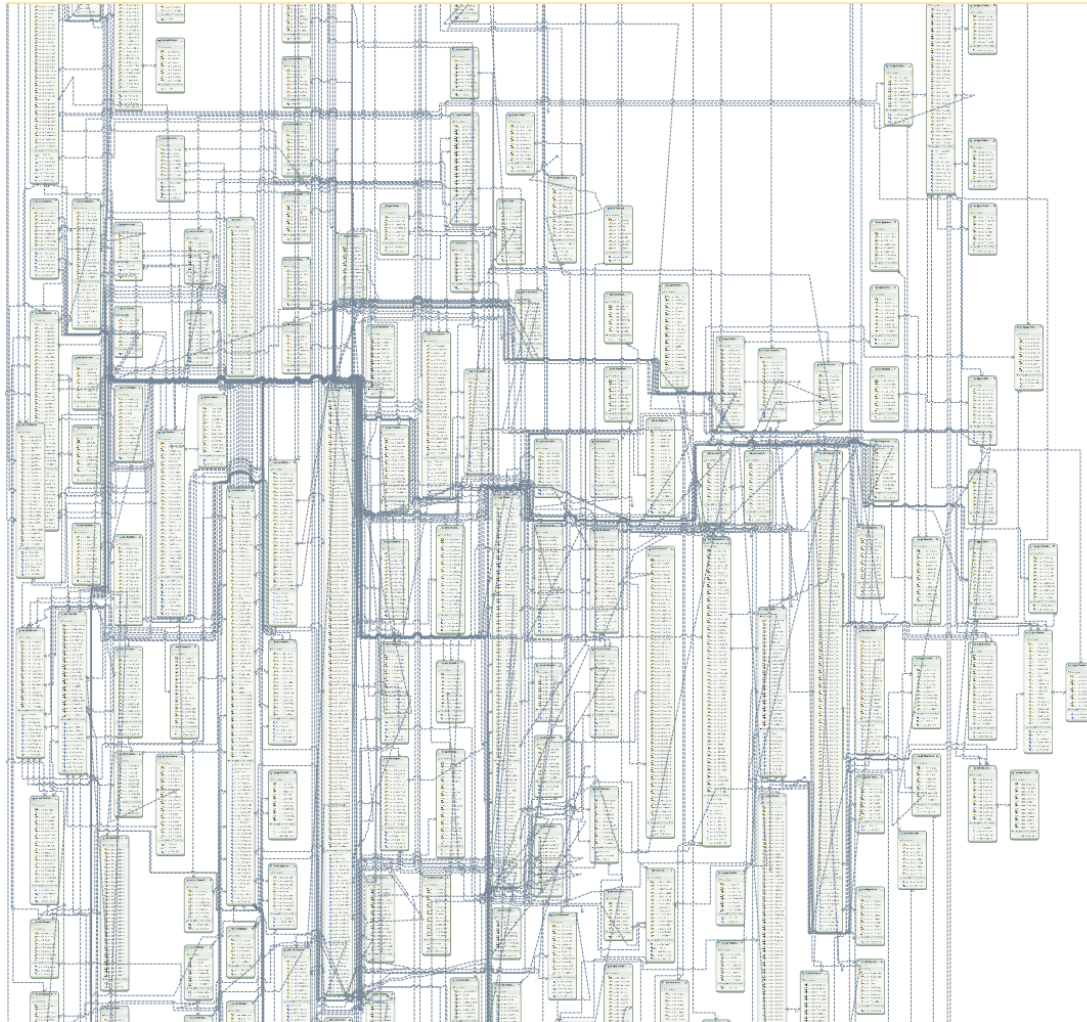


Requisitos, recopilación, comprensión, preparación de Datos

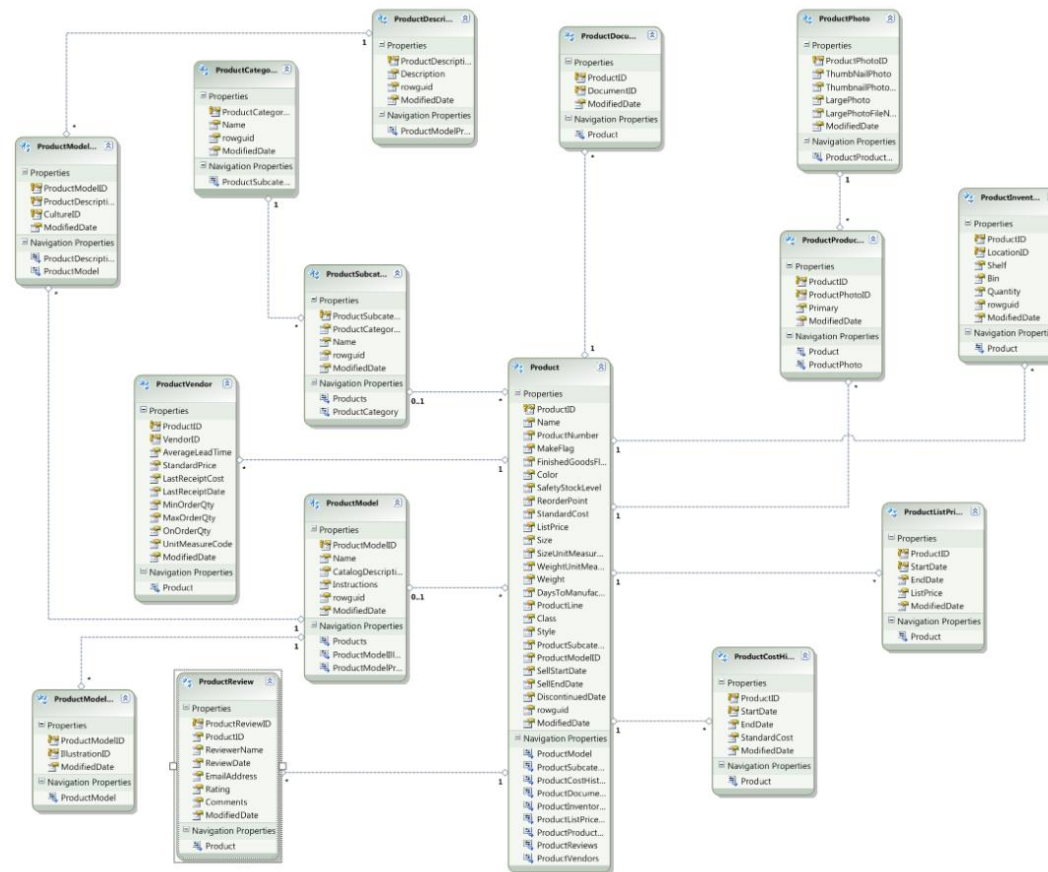




Modelo de Datos Complejo

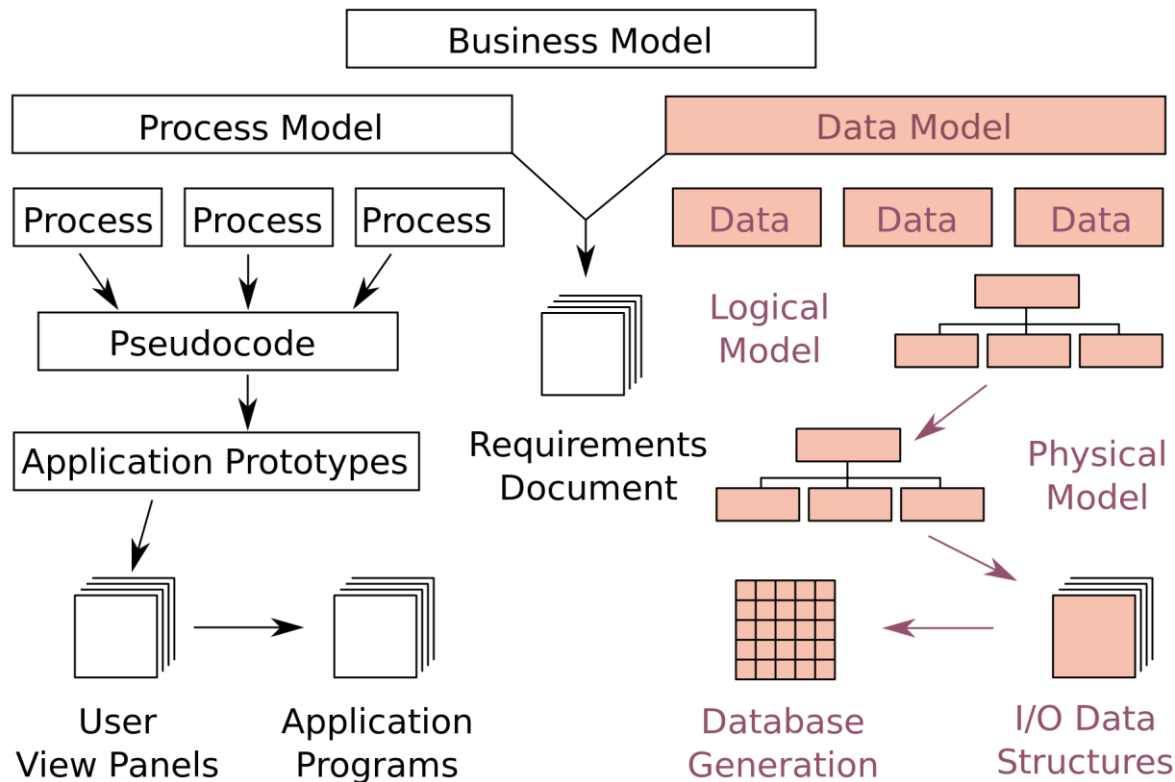


Modelo de Datos (I)

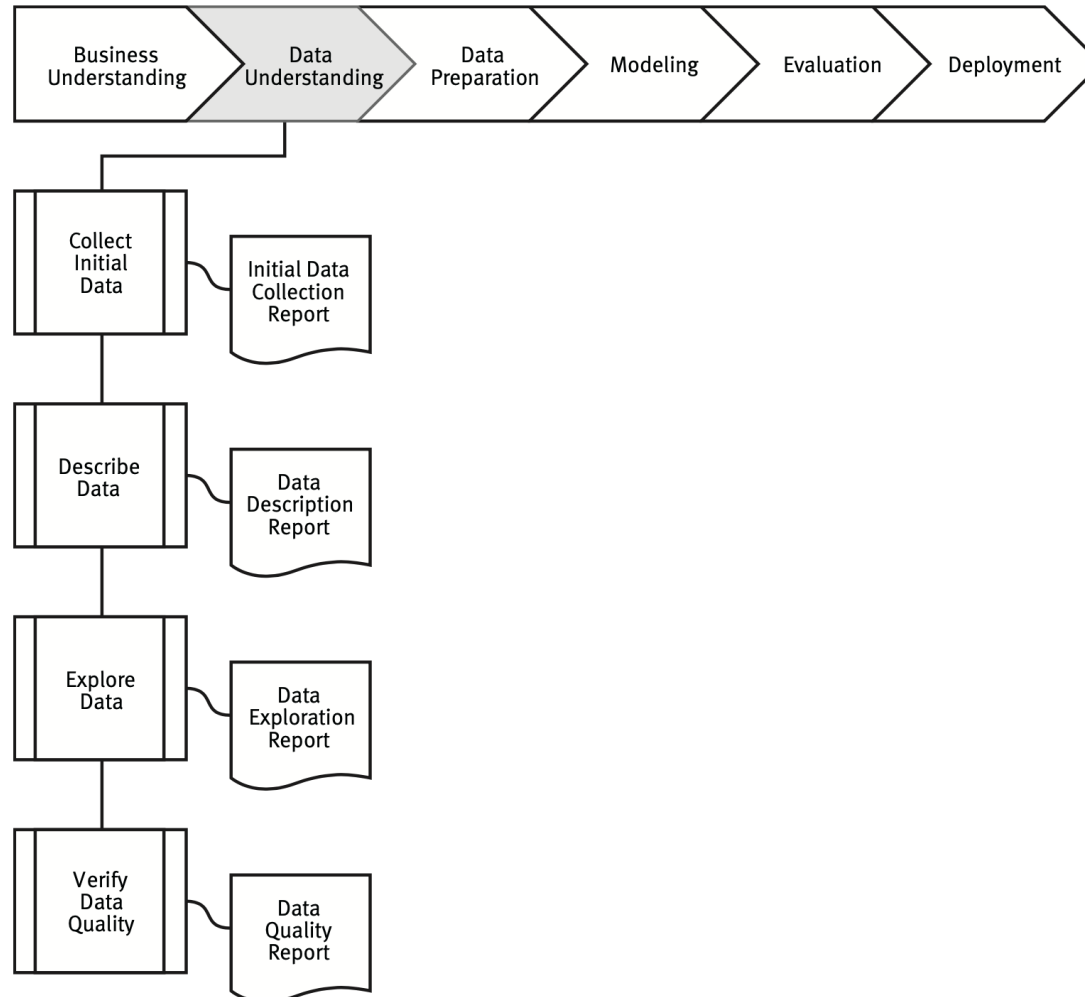


Modelo de Datos (II)

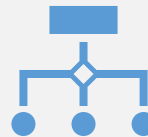
Business Model Integration



Fase: Entendimiento de datos



Entendimiento de los datos



IMPLICA ACCEDER A LOS DATOS Y EXPLORARLOS CON LA AYUDA DE TABLAS Y GRÁFICOS QUE SE PUEDEN ORGANIZAR.

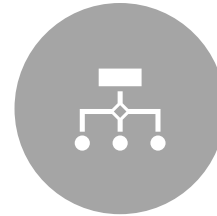


SE SUELEN UTILIZAR ESTADÍSTICAS DESCRIPTIVAS Y TÉCNICAS DE VISUALIZACIÓN PARA COMPRENDER EL CONTENIDO DE LOS DATOS, EVALUAR SU CALIDAD Y DESCUBRIR INSIGHTS INICIALES SOBRE ELLOS.

Entendimiento de los datos



Recopilación de
Datos



Descripción de los
datos



Explorar datos



Verificar la calidad
de los datos

Recopilación de Datos

Se reúnen los recursos de datos disponibles (estructurados, no estructurados y semiestructurados) y relevantes para el dominio del problema.

Al incorporar más datos, los modelos predictivos pueden representar mejor los eventos raros, como la incidencia de una enfermedad o un fallo del sistema.

La procedencia de los datos



Datos existentes. Incluye una amplia variedad de datos, como datos transaccionales, datos de encuesta, registros Web, etc. Tener en cuenta si los datos existentes son suficientes para adaptarse a las necesidades y objetivos a cumplir.



Datos adquiridos. ¿Cómo organización se utilizan datos adicionales, como datos demográficos? Si no se utiliza, hay que considerar si son necesarios.



Datos adicionales. Si las fuentes anteriores no satisfacen las necesidades, es posible que se necesite realizar encuestas o realizar seguimientos adicionales para servir de complemento a los repositorios de datos actuales.

Informe inicial de recolección de datos



Describa todos los diversos datos utilizados para el proyecto e incluya 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.



Cualquier evaluación de la calidad de los datos debe hacerse no solo de las fuentes de datos individuales sino también de los datos que resultan de la fusión de las fuentes de datos.



Debido a inconsistencias entre las fuentes, los datos combinados pueden presentar problemas que no existen en las fuentes de datos individuales.



Planificación de requerimientos de datos

- **Planificar** qué información se necesita (p. Ej., Solo para atributos dados o información adicional específica).
- **Comprobar** si toda la información necesaria (para resolver los objetivos de minería de datos) está realmente disponible.

Criterio de selección



Especifique los criterios de selección (por ejemplo, ¿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 debe usar un historial (por ejemplo, incluso si hay 18 meses de datos disponibles, solo se necesitarán 12 meses para el ejercicio).



Inserción de datos

¿Qué atributos (columnas) de la base de datos parecen más prometedores?

¿Qué atributos no parecen relevantes y se pueden excluir?

¿Existen datos suficientes para obtener conclusiones generales o realizar predicciones precisas?

¿Se dispone de atributos suficientes para el método de modelado?

Recuerde que algunos conocimientos sobre los datos pueden estar disponibles de fuentes no electrónicas (por ejemplo, de personas, texto impreso, etc.).

¿Se está fusionando varios orígenes de datos? En caso afirmativo, ¿existen áreas que puedan plantear problemas al fusionar?

Si los datos contienen entradas de texto libre, ¿debemos codificarlos para modelar o queremos agrupar entradas específicas?

¿Cómo se pueden adquirir los atributos faltantes?

¿Cómo podemos extraer mejor los datos?

Ejemplo de venta en línea: recopilación inicial de datos



Registros Web. Los registros de acceso brutos contienen toda la información de cómo los clientes navegan por el sitio Web. Es necesario eliminar referencias a archivos de imágenes y entradas no informativas en los registros Web como parte del proceso de preparación de datos.



Adquisición de datos. Si un cliente envía un pedido, se guarda toda la información relativa a ese pedido. Los pedidos de la base de datos de adquisiciones se deben asignar a las sesiones correspondientes en los registros Web.



Base de datos de productos. Los atributos de productos pueden ser de gran utilidad cuando determine productos “relacionados”. Es necesario asignar la información de productos a los pedidos correspondientes.



Base de datos de clientes. Esta base de datos contiene información adicional recopilada de clientes registrados. Los registros no son completos ya que muchos clientes no completan los cuestionarios. Es necesario asignar la información de los clientes a las adquisiciones y sesiones correspondientes en los registros Web.

Descripción de los datos



Existen muchas formas de describir datos, pero la mayoría de datos se centra en la cantidad y calidad de los datos.



Cantidad de datos. Los grandes conjuntos de datos pueden producir modelos más precisos, pero también pueden aumentar el tiempo de procesamiento.



Tipos de valores. Los datos pueden incluir una variedad de formatos, como numérico, categórico (cadena) o Booleano (verdadero/falso).



Esquemas de codificación. Con frecuencia, los valores de la base de datos son representaciones de características como género o tipo de producto.



Informe de descripción de datos



El formato de los datos



La cantidad de datos (por ejemplo, el número de registros y campos dentro de cada tabla)



La calidad de datos.



Las identidades de los campos



Cualquier otra característica de superficie que se haya descubierto.

Análisis volumétrico de datos.



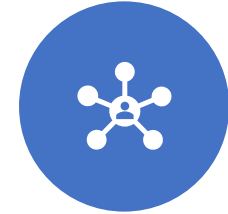
IDENTIFICAR DATOS Y
MÉTODO DE CAPTURA.



ACCEDER A FUENTES DE
DATOS.



UTILIZAR ANÁLISIS
ESTADÍSTICOS SI
CORRESPONDE.



TABLAS DE INFORMES Y
SUS RELACIONES.



VERIFICAR EL VOLUMEN
DE DATOS, NÚMERO DE
MÚLTIPLOS,
COMPLEJIDAD.



Tipos de atributos y valores



Verificar la accesibilidad y disponibilidad de atributos.



Verifique los tipos de atributos (numéricos, simbólicos, taxonómicos, etc.)



Comprobar rangos de valores de atributos.



Analizar correlaciones de atributos.



Comprender el significado de cada atributo y valor de atributo en términos comerciales.



Analizar estadísticas básicas y relacionar los resultados con su significado en términos comerciales.



Decidir si el atributo es relevante para el objetivo específico de minería de datos

Ejemplo de venta en línea: descripción de los datos



Existen multitud de registros y atributos para procesar en una aplicación de minería Web.



Aunque el área de negocio que realice el proyecto de minería de datos haya limitado el estudio inicial unos 30.000 clientes aproximadamente, que se hayan registrado en el sitio, aún quedan millones de registros en los registros Web.



La mayoría de tipos de valor de estos orígenes de datos son simbólicos, ya sean fechas y horas, accesos de páginas Web o respuestas a preguntas de opciones múltiples del cuestionario de registro.



Algunas de estas variables se utilizarán para crear nuevas variables numéricas, como el número de páginas Web visitadas y el tiempo que se ha permanecido en el sitio Web.

Ejemplo de venta en línea: descripción de los datos



Las pocas variables numéricas existentes en los conjuntos de datos incluyen el número de cada producto solicitado, la cantidad gastada durante una compra y las especificaciones de peso y dimensiones de la base de datos del producto.



Los esquemas de codificación de los diferentes orígenes de datos se solapan muy poco, porque los orígenes de datos contienen atributos muy diferentes. Las únicas variables que se solapan son “claves”, como las ID de clientes y códigos de productos.



Estas variables deben tener esquemas de codificación idénticos desde un origen de los datos a otro; de otro modo será imposible fundir los orígenes de datos.



Deberá realizar una preparación adicional de los datos para volver a codificar estos campos clave para fusionar.

Explorar datos



Tarea que aborda las preguntas de minería de datos que pueden cubrirse utilizando técnicas de consulta, visualización e informes.



Los análisis pueden abordar directamente los objetivos de minería de datos.

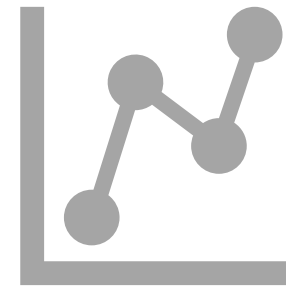


Pueden contribuir o refinar la descripción de los datos y los informes de calidad, y alimentar la transformación y otros pasos de preparación de datos necesarios antes de que pueda realizarse un análisis adicional.

Informe de exploración de datos (I)



Describe los resultados de esta tarea, incluidos los primeros hallazgos o hipótesis iniciales y su impacto en el resto del proyecto.



El informe también puede incluir gráficos y diagramas que indican características de datos o apuntan a subconjuntos de datos interesantes que merecen un examen más detallado.

Informe de exploración de datos (II)

¿Qué tipo de hipótesis sobre los datos ha formulado?

¿Qué atributos parecen ser prometedores de cara a futuros análisis?

¿Ha realizado exploraciones que revelen nuevas características de los datos?

¿En qué forma han cambiado estas exploraciones su hipótesis inicial?

¿Puede identificar subconjuntos concretos de datos para un uso posterior?

Vuelva a comprobar sus objetivos de minería de datos. ¿Esta exploración ha modificado sus objetivos?

Exploración de datos



ANALICE LAS PROPIEDADES DE
LOS ATRIBUTOS INTERESANTES
EN DETALLE (POR EJEMPLO,
ESTADÍSTICAS BÁSICAS,
SUBPOBLACIONES
INTERESANTES)



IDENTIFICAR LAS
CARACTERÍSTICAS DE LAS
SUBPOBLACIONES.



USAR GRÁFICOS DE
VISUALIZACIÓN, HISTOGRAMAS,
ETC. PARA REVELAR
INCONSISTENCIAS EN LOS
DATOS.

Suposiciones de forma para análisis futuros



Considerar y evaluar la información y los hallazgos en el informe de descripciones de datos.



Formar una 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. Una búsqueda "ciega" no es necesariamente inútil, pero es preferible una búsqueda más dirigida hacia los objetivos comerciales.



Realizar análisis básicos para verificar la hipótesis.

Verificar la calidad de los datos (I)

La mayoría de los datos contienen errores de codificación, valores perdidos u otro tipo de incoherencias que hacen que los análisis resulten difíciles en algunas ocasiones.



Una forma de evitar posibles problemas es realizar un análisis de calidad de los datos disponibles antes de proceder al modelado.



Verificar la calidad de los datos (II)

¿Están los datos completos (cubre todos los casos requeridos)?

¿Es correcto o contiene errores?

Si hay errores, ¿qué tan comunes son?

¿Faltan valores en los datos?

Si es así, ¿cómo están representados, ¿dónde ocurren y qué tan comunes son?

Verificación de calidad de datos



Los datos perdidos incluyen valores vacíos o codificados como sin respuesta (como null, ?, o 999).



Los errores de datos suelen ser errores tipográficos cometidos al introducir los datos.



Los errores de mediciones incluyen datos que se introducen correctamente, pero se basan en un esquema de mediciones incorrecto.



Las incoherencias de codificación suelen incluir unidades no estándar de medidas o valores incoherentes, como el uso de “M” y “masculino” para expresar el genero.



Los metadatos erróneos incluyen errores entre el significado aparente de un campo incluido en un nombre o definición de campo.

Revisar claves, atributos



VERIFIQUE LA COBERTURA (POR EJEMPLO, SI TODOS LOS VALORES POSIBLES ESTÁN REPRESENTADOS)



VERIFICAR KEYS



VERIFIQUE QUE LOS SIGNIFICADOS DE LOS ATRIBUTOS Y LOS VALORES CONTENIDOS COINCIDAN



IDENTIFICAR ATRIBUTOS FALTANTES Y CAMPOS EN BLANCO.



ESTABLECER EL SIGNIFICADO DE LOS DATOS FALTANTES



VERIFIQUE LOS ATRIBUTOS CON DIFERENTES VALORES QUE TIENEN SIGNIFICADOS SIMILARES (POR EJEMPLO, BAJO EN GRASAS, DIETA)



Calidad de datos en archivos planos

Si los datos se almacenan en archivos planos, 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 el número de campos en cada registro para ver si coinciden

Ejemplo de venta en línea: Exploración de datos

La verificación de la calidad de los datos se suele realizar durante el curso de los procesos de descripción y exploración. Algunos de los problemas detectados por el negocio suelen incluir:



Datos perdidos. Los datos perdidos conocidos incluyen cuestionarios sin responder por parte de algunos usuarios registrados. Sin la información extra que proporciona este cuestionario, estos clientes se pueden omitir en algunos de los siguientes modelos.



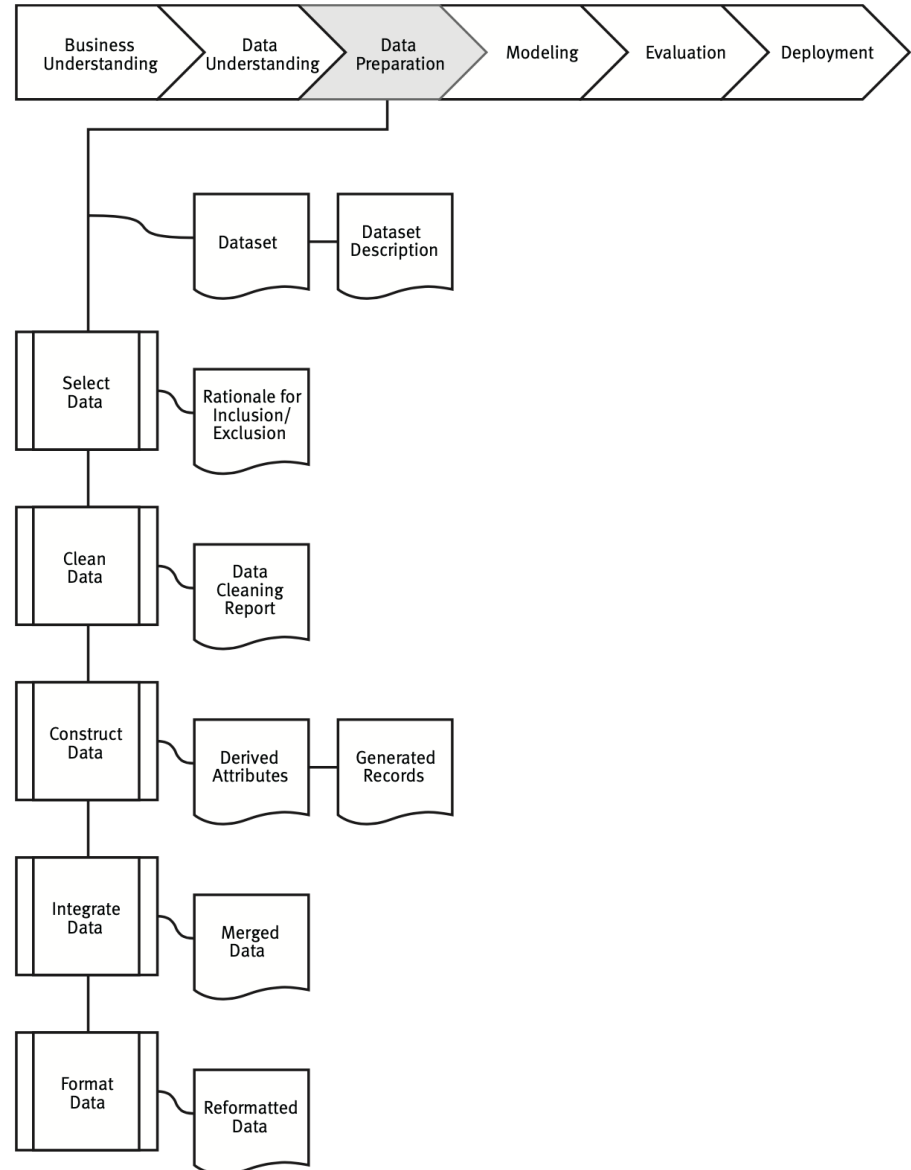
Errores de datos. La mayoría de los orígenes de datos se generan automáticamente, por lo que no es un problema grave. Los errores tipográficos de la base de datos de producto se pueden detectar durante el proceso de exploración.



Errores de mediciones. El origen principal de los errores de mediciones es el cuestionario. Si alguno de los elementos no está complementado correctamente, es posible que no proporcione la información que el negocio espera obtener. De nuevo, durante el proceso de exploración, es importante prestar una especial atención a los elementos que tienen una distribución inusual de las respuestas.



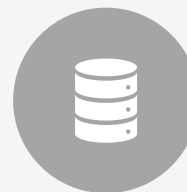
Fase: Preparación de datos



Preparación de datos (I)



Esta etapa abarca todas las actividades para construir el conjunto de datos que se utilizará en la subsiguiente etapa de modelado



Entre las actividades de preparación de datos están la limpieza de datos (tratar con valores no válidos o que faltan, eliminar duplicados y dar un formato adecuado), combinar datos de múltiples fuentes (archivos, tablas y plataformas) y transformar los datos en variables más útiles.

Preparación de datos (II)



Los científicos de datos utilizan un proceso llamado ingeniería de características (*feature engineering*) para crear variables explicativas adicionales, también conocidas como indicadores o características, a través de una combinación de conocimiento en el dominio y de variables estructuradas existentes.



La preparación de datos es uno de los aspectos más importantes y con frecuencia que más tiempo exige.



En muchos dominios, algunos pasos de la preparación de datos son comunes para problemas diferentes.

Preparación de datos (III)



La preparación de datos es uno de los aspectos más importantes y con frecuencia que más tiempo exige.



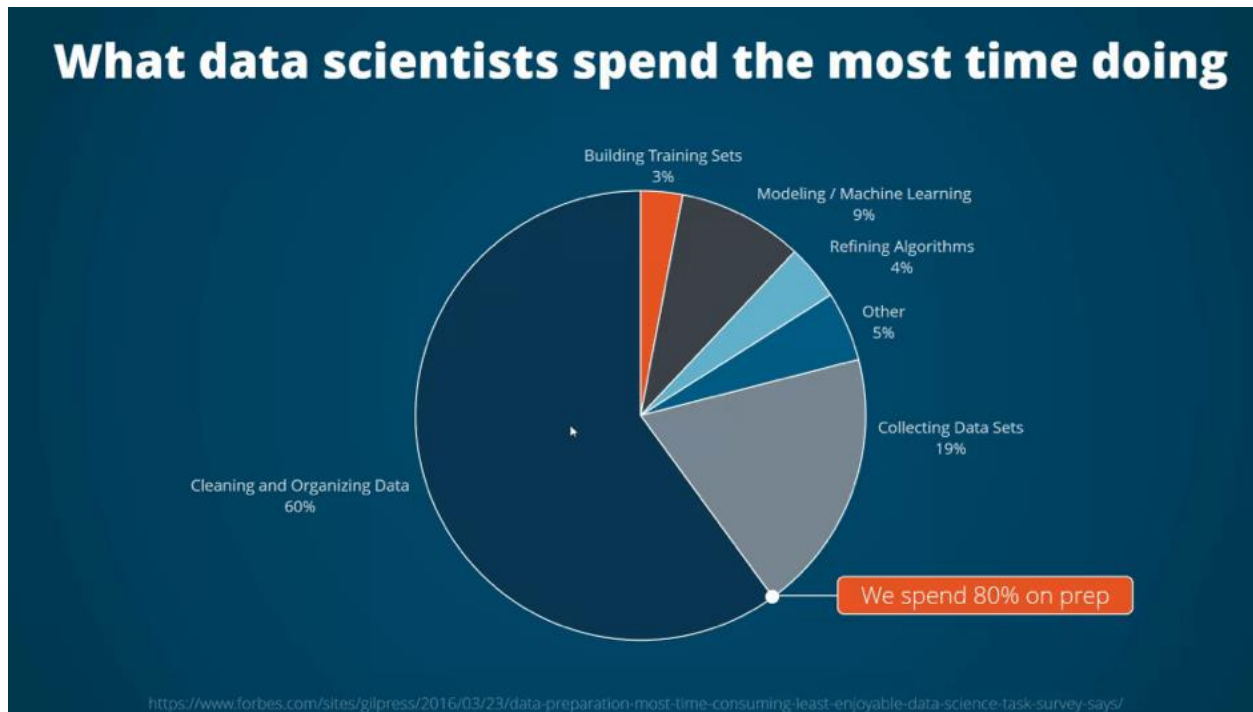
De hecho, se estima que la preparación de datos suele llevar el 50-80 % del tiempo y esfuerzo de un proyecto.



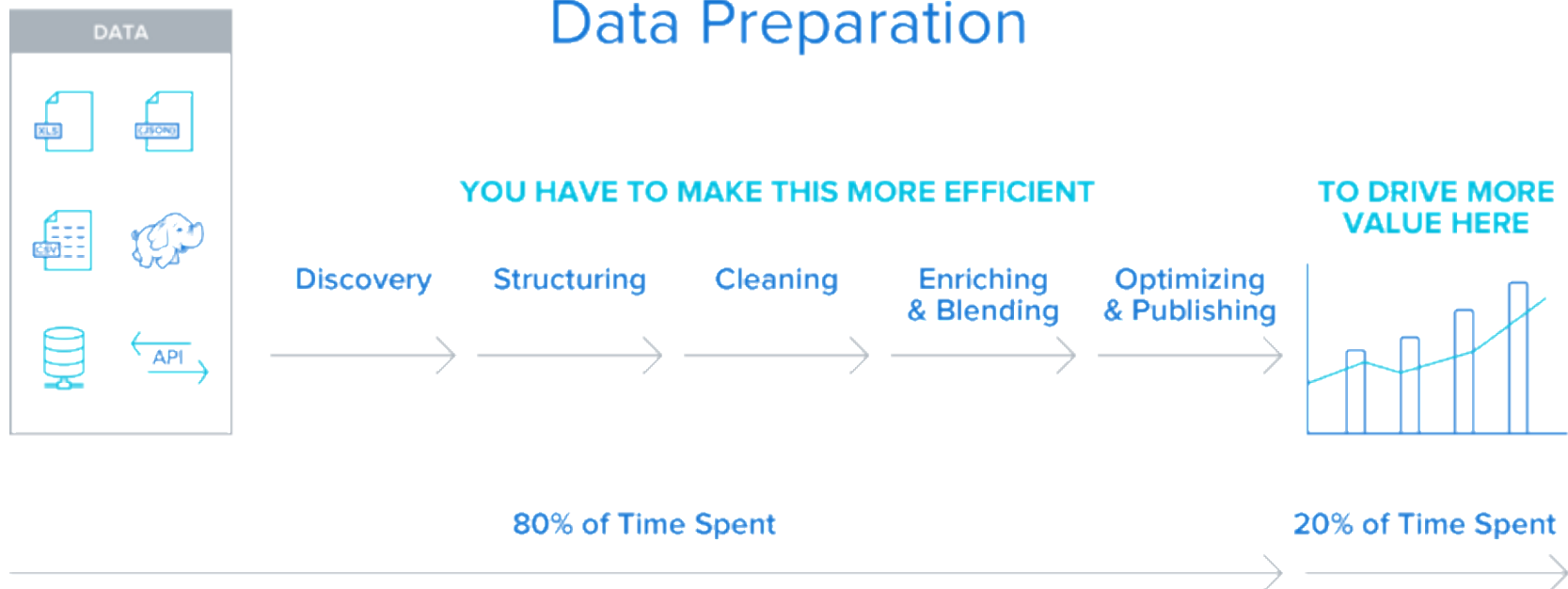
Dedicar los esfuerzos adecuados a las primeras fases de comprensión comercial y comprensión de datos puede reducir al mínimo los gastos indirectos relacionados, pero aún se deberá dedicar una buena cantidad de esfuerzo para preparar y empaquetar los datos.



Preparación de datos – Estadísticas (I)



Preparación de datos – Estadísticas (II)



Data Preparation is the process of cleaning, structuring and enriching raw data into a desired output for analysis.

Preparación de los datos



Selección de
datos



Limpieza de datos



Construcción de
nuevos datos



Integración de
datos



Formato de datos



Tareas en la preparación de datos



Fusión de conjuntos
y/o registros de datos.



Selección de una
muestra de un
subconjunto de datos.



Agregación de
registros.



Derivación de nuevos
atributos.



Clasificación de los
datos para el
modelado.



Eliminación o
sustitución de valores
en blanco o ausentes.



División en conjuntos
de datos de prueba y
entrenamiento .

Conjunto de datos

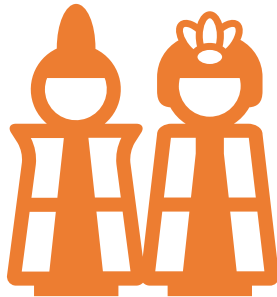


Dataset. Estos son los conjuntos de datos producidos por la fase de preparación de datos, utilizados para el modelado o para el trabajo de análisis principal del proyecto.



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

Selección de datos



Selección de elementos (filas) implica la toma de decisiones como las cuentas, productos o clientes que se van a incluir.



Selección de atributos o características (columnas) implica la toma de decisiones sobre el uso de características como la cantidad de las transacciones o los ingresos por hogar.

Ejemplo de venta en línea: selección de datos



Selección de elementos. El estudio inicial se limitará a los (aproximadamente) 30.000 clientes registrados en el sitio, por lo que es necesario configurar los filtros para que excluyan las compras y los registros Web de clientes sin registrar. Otros filtros se deben configurar para eliminar llamadas a archivos de imágenes y otras entradas no informativas de los registros Web.



Selección de atributos. La base de datos de adquisiciones contendrá información confidencial de los clientes de la empresa, por lo que es importante filtrar los atributos como nombre del cliente, dirección, número de teléfono y números de tarjeta de crédito.

Inclusión o exclusión de datos



Cuestiones que debe tener en cuenta



¿Existe un atributo relacionado con sus objetivos de minería de datos?



¿La calidad de un conjunto o atributo de datos concreto excluye la validez de los resultados?
¿Puede recuperar estos datos?



¿Existen limitaciones acerca del uso de campos concretos como género o raza?

Limpieza de datos

La limpieza de datos implica observar más de cerca los problemas en los datos que ha seleccionado incluir en el análisis.

Problema de datos	Solución posible
Datos perdidos	Excluya las filas o características, también puede complementarlas con un valor estimado.
Errores de datos	Utilice recursos lógicos para descubrir errores manuales y corríjalos. O, excluya las características.
Incoherencias de codificación	Decida un esquema de codificación simple y convierta y sustituya los valores.
Metadatos ausentes o erróneos	Examine manualmente los campos sospechosos y compruebe el significado correcto.

El informe de calidad de datos preparado durante la fase de comprensión de datos contiene detalles sobre los tipos de problemas concretos de sus datos.



Ejemplo de venta en línea: limpieza de los datos (I)

Datos perdidos. Es posible que los clientes que no han completado el cuestionario en línea se tengan que omitir de los modelos posteriores.

Es necesario volver a pedir a estos clientes que completen el cuestionario, pero esta solución requiere dedicar tiempo y dinero que es posible que el negocio no pueda invertir.

Lo que el negocio puede hacer es modelar las diferencias de compras entre los clientes que responden y los que no responden el cuestionario.

Si estos dos conjuntos de clientes tienen hábitos de compra similares, los cuestionarios que faltan son menos preocupantes.

Ejemplo de venta en línea: limpieza de los datos (II)



Errores de datos. Los errores detectados durante el proceso de exploración se pueden corregir en esta fase.



La mayoría de las veces, sin embargo, la entrada correcta de datos se realiza en el sitio Web antes de que un cliente envíe una página a la base de datos de back-end.



Ejemplo de venta en línea: limpieza de los datos (III)

Errores de mediciones. Los elementos incorrectos del cuestionario pueden afectar en gran medida a la calidad de los datos.

Al igual que los cuestionarios perdidos, se trata de un problema difícil, porque es posible que no se disponga del tiempo o recursos disponibles para recopilar las respuestas a una nueva pregunta.

Para elementos problemáticos, la mejor solución puede ser volver al proceso de selección y filtrar estos elementos de análisis posteriores.

Escritura de un informe de limpieza de datos



Es una excelente idea considerar las siguientes cuestiones cuando genere el informe:



¿Qué tipos de ruido se han producido en los datos?



¿Qué métodos utiliza para eliminar el ruido? ¿Qué técnicas han demostrado ser eficaces?



¿Existen casos o atributos que no se pueden recuperar? Asegúrese de registrar los datos que se han excluido por causas del ruido.

Construcción de nuevos datos

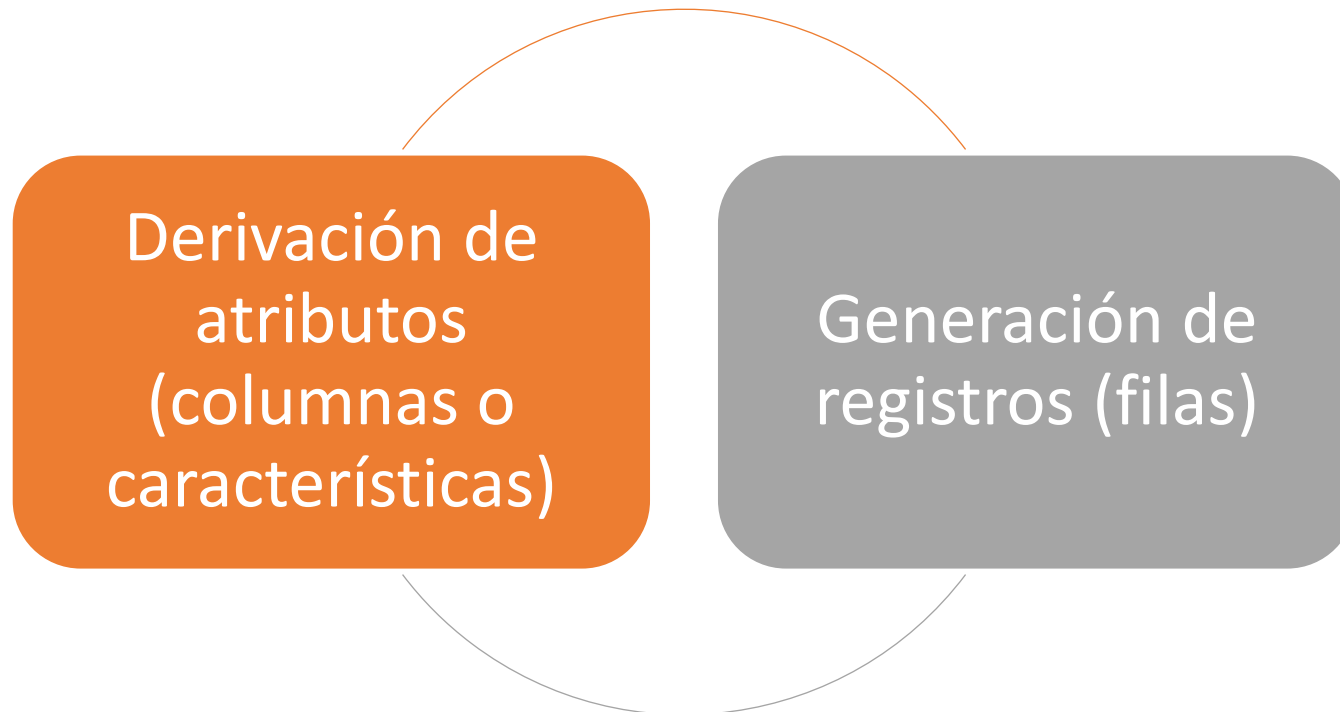
Con frecuencia, necesitará construir nuevos datos.



Por ejemplo, puede ser de gran utilidad crear una nueva columna con la adquisición de una garantía ampliada en cada transacción.



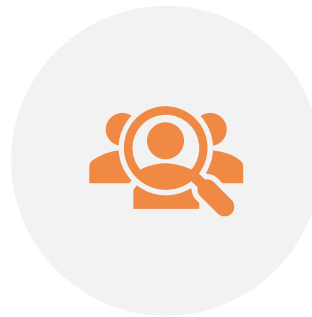
Existen dos formas de construir nuevos datos



Ejemplo de venta en línea: construcción de los datos (I)



EL PROCESAMIENTO DE REGISTROS WEB PUEDE CREAR MÚLTIPLES REGISTROS NUEVOS.



EN LOS CASOS REGISTRADOS, EL NEGOCIO PUEDE PREFERIR CREAR MARCAS DE TIEMPO, IDENTIFICAR VISITANTES Y SESIONES Y REGISTRAR LA PÁGINA A LA QUE SE HA ACCEDIDO Y EL TIPO DE ACTIVIDAD DEL EVENTO.



ALGUNAS DE ESTAS VARIABLES SE UTILIZARÁN PARA CREAR MÁS ATRIBUTOS, COMO LOS TIEMPOS ENTRE EVENTOS EN UNA SESIÓN.

Ejemplo de venta en línea: construcción de los datos (II)



Se pueden crear más atributos como resultado de una fusión u otra reestructuración de los datos.



Por ejemplo, si los registros Web de evento por fila se “acumulan” de forma que cada fila sea una sesión, se crearán nuevos atributos que registran el número total de acciones, el tiempo total empleado y el número total de compras realizadas durante la sesión.

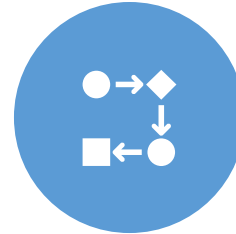


Si los registros Web se fusionan con la base de datos del cliente de forma que cada fila es un nuevo cliente, se crearán los nuevos atributos que registran el número de sesiones, el número total de acciones, el tiempo total empleado y el número de compras totales realizadas por cada cliente.

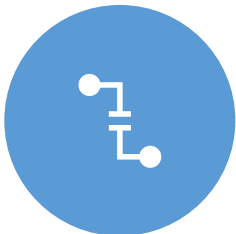
Derivación de atributos



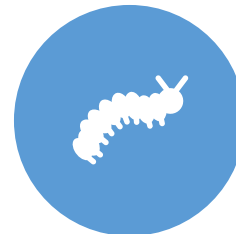
Tenga en cuenta los requisitos de datos de modelado cuando derive atributos. ¿El algoritmo de modelado espera un tipo de datos concreto, como datos numéricos? En caso afirmativo, realice las transformaciones necesarias.



¿Necesita normalizar los datos antes de proceder con el modelado?



¿Se pueden construir los atributos que faltan mediante la agregación, media o inducción?



En función de sus conocimientos, ¿existen hechos importantes (como la cantidad de tiempo en el sitio Web) que se puedan derivar de los campos existentes?



Integración de datos



Métodos básicos para integrar los datos

LA **FUSIÓN** DE DATOS IMPLICA UNIR DOS CONJUNTOS DE DATOS CON REGISTROS SIMILARES, PERO CON ATRIBUTOS DIFERENTES. LOS DATOS SE FUSIONAN UTILIZANDO EL MISMO IDENTIFICADOR CLAVE EN CADA REGISTRO (COMO EL ID DE USUARIO). LOS DATOS RESULTANTES AUMENTAN LAS COLUMNAS O LAS CARACTERÍSTICAS.

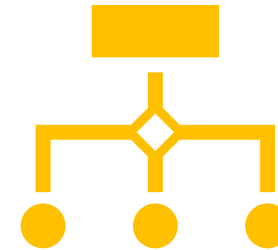


LA **ADICIÓN** DE DATOS IMPLICA INTEGRAR DOS O MÁS CONJUNTOS DE DATOS CON ATRIBUTOS SIMILARES, PERO CON REGISTROS DIFERENTES. LOS DATOS SE INTEGRAN EN FUNCIÓN DE LOS CAMPOS SIMILARES (COMO EL NOMBRE DE PRODUCTO O LA LONGITUD DEL CONTRATO).

Ejemplo de venta en línea: integración de datos (I)



Adición de atributos de clientes y productos a datos de eventos. Para modelar eventos de registros Web que utilicen atributos de otras bases de datos, cualquier ID de cliente, número de producto y número de orden de compra asociado con cada evento se debe identificar correctamente y los atributos correspondientes se deben fusionar con los registros Web procesados.



Tener en cuenta que el archivo replica la información del cliente y del producto cada vez que un cliente o producto se asocia con un evento.

Ejemplo de venta en línea: integración de datos (II)



Adición de información de compra y registro Web a los datos del cliente. Para modelar el valor de un cliente, su información de compras y de sesión se debe extraer de las bases de datos adecuadas, totalizadas y fusionadas con la base de datos de clientes.

Este método implica la creación de nuevos atributos, tal y como se explica en el proceso de construcción de datos.

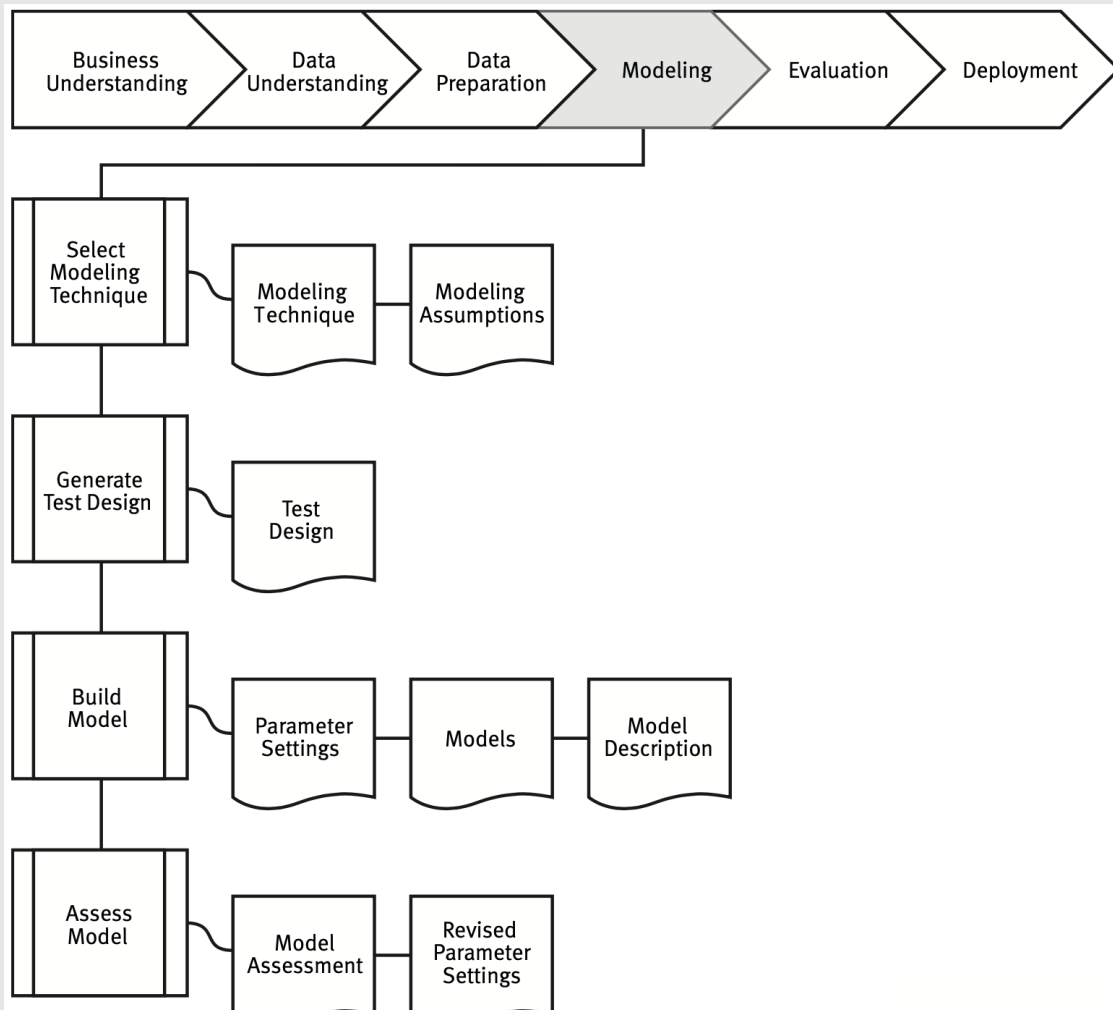
Formato de datos



Como paso final antes de la construcción del modelo, es muy útil comprobar si algunas técnicas requieren aplicar un formato concreto o la clasificación de los datos.



Por ejemplo, no es extraño que un algoritmo de secuencia requiera que los datos estén clasificados de forma previa antes de ejecutar el modelo. Incluso si el modelo puede ejecutar la clasificación de forma automática.



Fase:
Modelado

Modelado

Selección de técnicas de modelado analítico

Generación de un diseño de comprobación

Construcción de modelos analíticos

Evaluación del modelo analítico

Modelado



Los Data Scientist ejecutan varios modelos utilizando los parámetros por defecto y ajustan los parámetros o vuelven a la fase de preparación de datos para las manipulaciones necesarias por su modelo.



El modelado se suele ejecutar en múltiples iteraciones.



Es extraño que las cuestiones relativas a la minería de datos de una empresa se solucionen satisfactoriamente con un modelo y ejecución únicos.



Selección de técnicas de modelado analítico



Como primer paso en el modelado, es seleccionar la técnica de modelado real que se utilizará.



Esta tarea se refiere a la técnica de modelado específica, por ejemplo, la construcción del árbol de decisión, o la generación de redes neuronales con propagación hacia atrás.



Si se aplican varias técnicas, realice esta tarea por separado para cada técnica.



Selección de técnicas de modelado analítico

- La determinación del modelado más adecuado se basará en las siguientes consideraciones:
- **Los tipos de datos disponibles para la minería.** Por ejemplo, ¿los campos de interés son categóricos (simbólicos)?
- **Sus objetivos del análisis.** ¿Sólo quiere tener un mejor conocimiento de los almacenes de datos transaccionales y descubrir patrones de compras interesantes? ¿Necesita producir una puntuación indicando, por ejemplo, las posibilidades de impago de un préstamo a un estudiante?
- **Requisitos específicos de modelado.** ¿Necesita el modelo un tipo o un tamaño de datos concreto? ¿Necesita un modelo con unos resultados fácilmente presentables?



Ejemplo de venta en línea: técnicas de modelado

- **Recomendaciones mejoradas.** De forma simple, recopila los pedidos de compra agrupadas para determinar los productos que se adquieren conjuntamente con más frecuencia. Se pueden añadir datos de clientes e incluso registros de visita, para obtener unos resultados más completos.



Ejemplo de venta en línea: técnicas de modelado

- **Navegación mejorada por el sitio.** Ahora se centrará en identificar las páginas que se utilizan con más frecuencia pero que requieren que el usuario realice varias operaciones para llegar a ellas. Implica aplicar un algoritmo de secuencia a los registros Web para generar las “rutas únicas” que los clientes utilizan en la Web y busca específicamente sesiones con multitud de visitas sin (o antes) de realizar una acción.

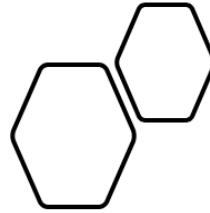


Selección de las técnicas de modelado correctas

- ¿Requiere el modelo que los datos se dividan en conjuntos de entrenamiento y prueba?
- ¿Dispone de datos suficientes para producir resultados fiables para un modelo concreto?
- ¿Requiere el modelo un cierto nivel de calidad de datos? ¿Puede alcanzar este nivel con los datos que dispone?
- ¿Son sus datos el tipo correcto para un modelo concreto? En caso contrario, ¿puede realizar las conversiones necesarias utilizando nodos de manipulación de datos?



Modelado de supuestos



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



Generación de un diseño de comprobación

- Antes de construir un modelo, se necesita generar un procedimiento o mecanismo para probar la calidad y validez del modelo.
- Por ejemplo, en tareas supervisadas de minería de datos como la clasificación, es común usar tasas de error como medidas de calidad para los modelos de minería de datos.
- Generalmente se separa el conjunto de datos en trenes y conjuntos de pruebas, se construye el modelo en el conjunto de trenes y estimamos su calidad en el conjunto de pruebas separado.

Prueba de Diseño (I)

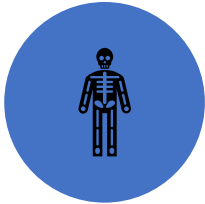


Describir el plan previsto para la capacitación, las pruebas y la evaluación de los modelos.

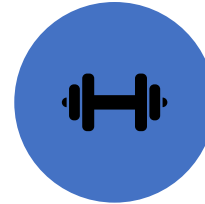


Un componente principal del plan es determinar cómo dividir el conjunto de datos disponible en conjuntos de datos de capacitación, prueba y validación.

Prueba de Diseño (II)



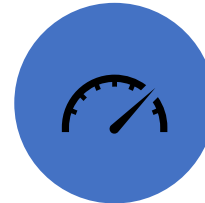
¿Qué datos se utilizarán para comprobar los modelos?



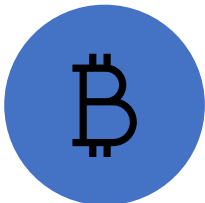
¿Ha particionado los datos en conjuntos de entrenamiento/prueba?



¿Cómo puede medir el rendimiento de modelos supervisados?



¿Cómo puede medir el rendimiento de modelos sin supervisar?



¿Cuántas veces piensa volver a ejecutar un modelo con los valores ajustados antes de intentar otro tipo de modelo?



Ejemplo de
venta en línea:
diseño de
comprobación

- **Recomendaciones mejoradas.** Hasta que se presenten las recomendaciones mejoradas a los clientes activos, no existe un método puramente objetivo de evaluarlas. Sin embargo, el negocio puede exigir que las reglas que generen las recomendaciones sean lo suficientemente simples para que tengan sentido desde una perspectiva comercial. Del mismo modo, las reglas deben ser lo suficientemente complejas para generar recomendaciones diferentes para clientes y sesiones diferentes.



Ejemplo de
venta en línea:
diseño de
comprobación

- **Navegación mejorada por el sitio.** Conociendo las páginas a las que acceden los clientes en el sitio Web, el negocio puede evaluar de forma objetiva el diseño del sitio actualizado en términos de facilidad de uso a las páginas más importantes. Sin embargo, al igual que con las recomendaciones, es difícil evaluar de forma anticipada cómo se ajustarán los usuarios al sitio reorganizado. Si lo permiten los plazos y el presupuesto, se pueden realizar comprobaciones de la facilidad de uso.



Generación de los modelos analíticos



ES NECESARIO
CONSIDERAR TIEMPO
PARA EXPERIMENTAR CON
DIFERENTES MODELOS
ANTES DE LLEGAR A
CONCLUSIONES
DEFINITIVAS.



LA MAYORÍA DE DATA
SCIENTIST SUELEN
GENERAR VARIOS
MODELOS Y COMPARAR
LOS RESULTADOS ANTES
DE APLICARLOS O
INTEGRARLOS.



Generación de los modelos analíticos

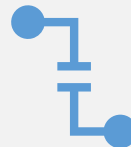
- Al final del proceso de generación de modelos se dispondrá de tres tipos de información que pueden utilizarse en la toma de decisiones del análisis:
- **Configuración de parámetros** incluye las notas que ha tomado sobre los parámetros que producen los mejores resultados.
- Los **modelos** reales producidos.
- **Descripciones de resultados de modelos**, incluyendo problemas de datos y rendimiento que hayan ocurrido durante la ejecución del modelo y exploración de los resultados.



Configuración de parámetros



La mayoría de técnicas de modelado tienen diferentes parámetros o configuraciones que se pueden ajustar para controlar el proceso de modelado.



Por ejemplo, los árboles de decisión se pueden controlar ajustando la profundidad del árbol, divisiones y otros ajustes.

Descripción de modelo

Quando se examinan los resultados de un modelo, asegurarse de tomar notas del proceso de modelado.

¿Puede llegar a conclusiones significativas a partir de este modelo?

¿Revela este modelo nuevas oportunidades o patrones alternativos?

¿El modelo presenta problemas de ejecución? ¿Fue razonable el tiempo de procesamiento?

¿El modelo presenta problemas de calidad de datos, como un alto número de valores perdidos?

¿Existen incoherencias de cálculos que se deben mencionar?



Evaluación del modelo



El data scientist interpreta los modelos de acuerdo con su conocimiento de dominio, los criterios de éxito de minería de datos y el diseño de prueba deseado.



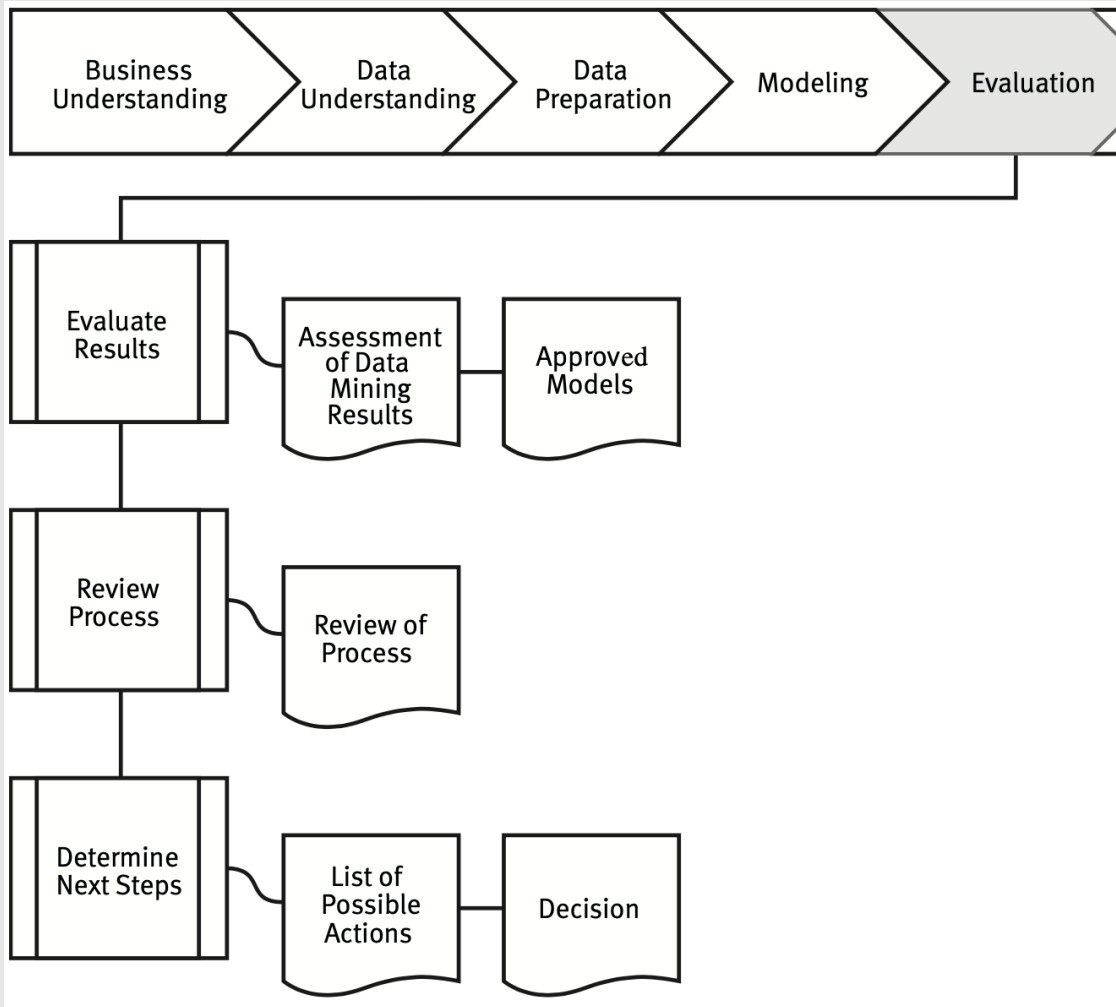
El científico de datos juzga técnicamente el éxito de la aplicación de técnicas de modelado y descubrimiento



El científico de datos se pone en contacto con analistas comerciales y expertos en dominios para analizar los resultados de la minería de datos en el contexto comercial.



Esta tarea solo considera modelos, mientras que la fase de evaluación también tiene en cuenta todos los demás resultados que se produjeron en el curso del proyecto.



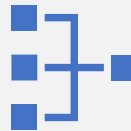
Fase:
Evaluación



Evaluación



Evaluación de los resultados



Proceso de revisión



Determinación de los pasos siguientes

Evaluación (I)



Se habrá determinado, en la fase de modelado, que los modelos son técnicamente correctos y efectivos en función de los criterios de rendimiento de minería de datos que se han definido previamente.



Se deben evaluar los resultados de los esfuerzos utilizando los criterios de rendimiento comercial establecidos en el inicio del proyecto.

Evaluación (II)



En este punto se habrá completado la mayor parte del proyecto de Analytics.



En la fase de modelado habrá determinado, que los modelos son técnicamente correctos y efectivos en función de los criterios de rendimiento que ha definido previamente.



Este paso evalúa el grado en que el modelo cumple con los objetivos comerciales y busca determinar si hay alguna razón comercial por la cual este modelo es deficiente.

Evaluación de los resultados

- En esta etapa, formalizará su evaluación en función de si los resultados del proyecto cumplen los criterios del rendimiento comercial.
- También se evalúan todos los demás hallazgos que no están necesariamente relacionados con los objetivos comerciales originales, pero que también pueden revelar desafíos, información o sugerencias adicionales para futuras direcciones.
- Este paso requiere una clara comprensión de los objetivos comerciales, por lo que debe estar seguro de incluir factores de toma de decisiones en la evaluación del proyecto.



Ejemplo de venta en línea: evaluación de resultados

- Los resultados globales de la primera experiencia del negocio son muy fáciles de comunicar desde una perspectiva comercial: el estudio refleja recomendaciones de mejora de producto y un diseño mejorado del sitio.
- El diseño mejorado del sitio se basa en la secuencias de navegación del cliente, que muestran las funciones del sitio que los clientes desean pero que requieren varios pasos.

Ejemplo de venta en línea: evaluación de resultados



- La prueba de que las recomendaciones de producto son de mejora es más difícil de comunicar, porque las reglas de decisión se pueden complicar.
- Para producir el informe final, los analistas intentarán identificar algunas tendencias generales en los conjuntos de reglas que se pueden explicar con mayor facilidad.



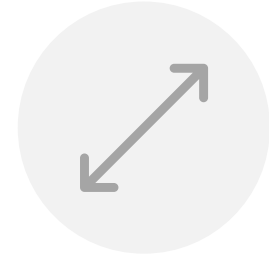
Ejemplo de
venta en
línea:
evaluación
de
resultados

- **Ordenación de los modelos.**
Como varios de los modelos iniciales parecían tener sentido comercial, la ordenación del grupo se basa en criterios estadísticos, fáciles de interpretar y de gran diversidad. Además, el modelo ofreció diferentes recomendaciones para diferentes situaciones.

Ejemplo de venta en línea: evaluación de resultados



NUEVAS CUESTIONES. LA CUESTIÓN MÁS IMPORTANTE QUE SURGE DEL ESTUDIO ES, ¿CÓMO PUEDE EL NEGOCIO TENER UN MAYOR CONOCIMIENTO DE SUS CLIENTES?



LA INFORMACIÓN EN LA BASE DE DATOS DE CLIENTES DESARROLLA UN IMPORTANTE PAPEL EN LA FORMACIÓN DE CONGLOMERADOS DE RECOMENDACIONES.



MIENTRAS EXISTEN REGLAS ESPECIALES PARA REALIZAR RECOMENDACIONES A LOS CLIENTES CUYA INFORMACIÓN FALTA, LAS RECOMENDACIONES SON MÁS GENERALES EN COMPARACIÓN CON LAS RECOMENDACIONES A LOS CLIENTES REGISTRADOS.

Proceso de revisión



En este punto, los modelos resultantes parecen ser satisfactorios y 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.



Esta revisión también cubre problemas de garantía de calidad, por ejemplo: ¿Creamos correctamente el modelo? ¿Utilizamos solo los atributos que tenemos permitido usar y que están disponibles para futuros análisis?

Determinación de los pasos siguientes



Dependiendo de los resultados de la evaluación y la revisión del proceso, el equipo del proyecto decide cómo proceder.



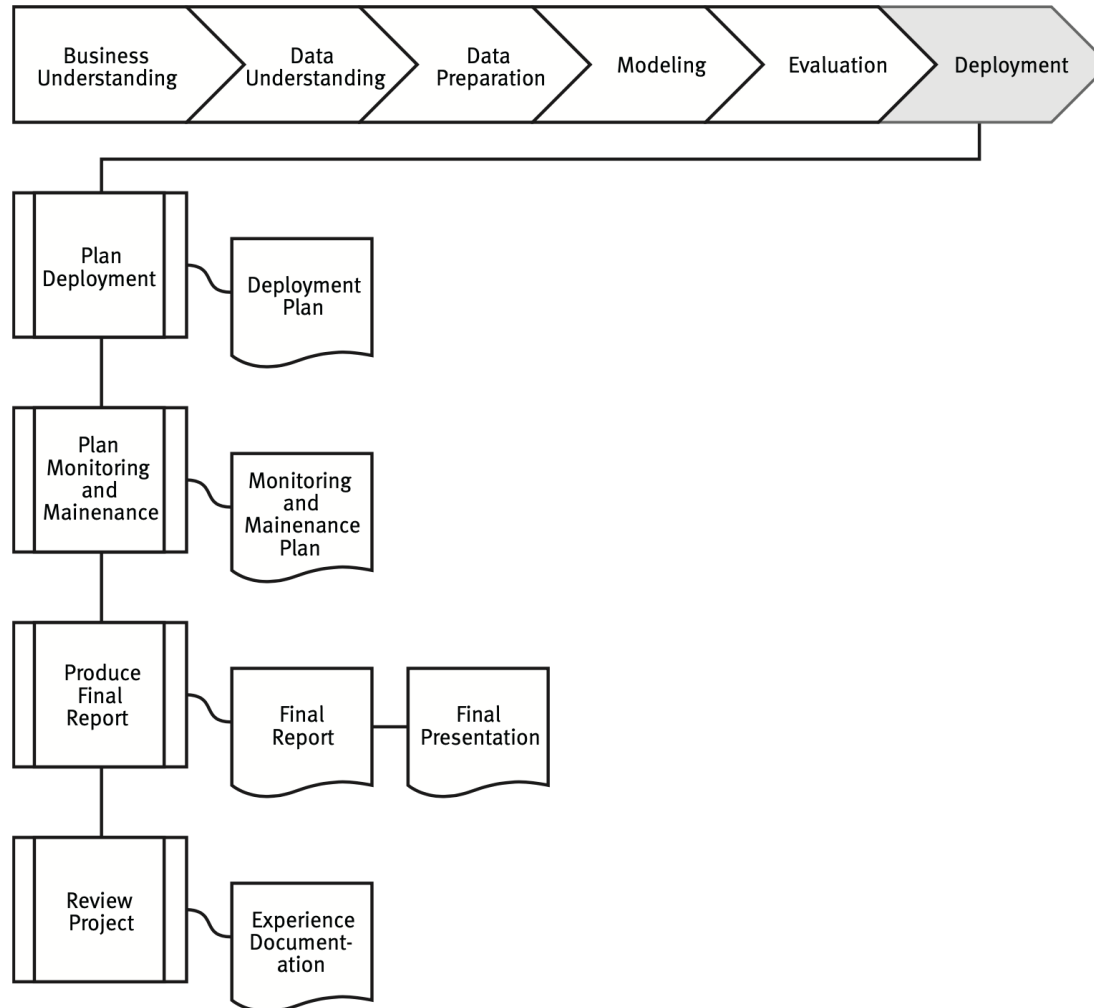
El equipo decide si finalizar este proyecto y continuar con la implementación, iniciar nuevas iteraciones o configurar nuevos proyectos de minería de datos.



Esta tarea incluye análisis de los recursos y el presupuesto restantes, que pueden influir en las decisiones.



Fase: Distribución



Distribución



Planificación de distribución



Planificación del control y del mantenimiento



Creación de un informe final



Revisión de Proyecto

Planificación de distribución



El primer paso es resumir los resultados; modelos y descubrimientos. Este método le ayudará a determinar los modelos que se pueden integrar en sus sistemas de base de datos y los descubrimientos que se presentarán a sus colegas.

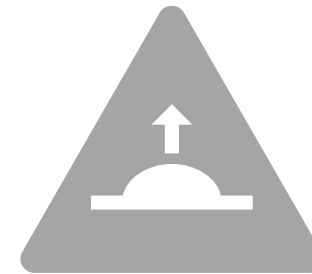


En cada modelo distribuible, cree una planificación paso a paso para la distribución e integración con sus sistemas. Registre los detalles técnicos como requisitos de base de datos para los resultados del modelo. Por ejemplo, es posible que su sistema requiera que los resultados del modelado se distribuyan en formato delimitado por tabulaciones.

Planificación del control y del mantenimiento



La supervisión y el mantenimiento son cuestiones importantes si el resultado de la minería de datos se convierte en parte del negocio diario y su entorno.



La preparación cuidadosa de una estrategia de mantenimiento ayuda a evitar períodos innecesariamente largos de uso incorrecto de los resultados de la minería de datos.

Planificación del control y del mantenimiento



En una distribución e integración completa de los resultados de modelado, el trabajo de analytics puede ser continuado. Por ejemplo, si un modelo se distribuye para pronosticar las consecuencias de las compras en línea, es probable que este modelo se tenga que evaluar periódicamente para asegurar su eficacia y realizar mejoras continuas.



Del mismo modo, un modelo distribuido para aumentar la retención de los clientes más importantes se deberá modificar una vez se ha alcanzado un nivel concreto de retención.



El modelo se puede modificar y reutilizar para retener clientes de un nivel inferior, pero que siguen teniendo un nivel de rentabilidad en la pirámide de valores.

Creación de un informe final



La escritura de un informe final no sólo resuelve los cabos sueltos de la documentación previa, sino que también se utiliza para comunicar los resultados.



Es importante presentar los resultados a las diferentes personas relacionadas con los resultados.



Se pueden incluir a los administradores técnicos, que son responsables de la aplicación de los resultados de modelado, así como el departamento de marketing y gestión, encargado de tomar las decisiones en función de los resultados obtenidos.

Revisión de Proyecto (I)



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



Resumir la importante experiencia obtenida durante el proyecto.



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 también cubre cualquier informe que haya sido escrito por miembros individuales del proyecto durante las fases anteriores del proyecto.

Revisión de Proyecto (II)



¿Cuál es su impresión global del proyecto?



¿Qué conocimientos ha adquirido durante el proceso en general y los datos disponibles?



¿Qué partes del proyecto han funcionado correctamente?



¿Dónde han surgido las dificultades?



¿Existe algún tipo de información que le podría haber evitado confusiones?



Ejemplo
de venta
en línea:
revisión
final

- **Entrevistas a miembros del proyecto.** El negocio encuentra que los miembros del proyecto más íntimamente relacionados con el estudio de inicio a fin son en su mayoría entusiastas con los resultados y esperan poder implicarse en proyectos futuros. El grupo de la base de datos tiene un optimismo contenido y aunque que aprecian la utilidad del estudio, señalan la carga añadida que suponen los recursos de la base de datos. Durante el estudio, se dispuso de la ayuda de un asesor, pero en el futuro se necesitará un empleado dedicado al mantenimiento de la base de datos a medida que se amplíe el proyecto.



Ejemplo de
venta en
línea:
revisión
final

- **Entrevistas de clientes.** Los comentarios de los clientes han sido muy positivos hasta el momento. Uno de los temas que no tuvo una buena previsión fue el impacto de las modificaciones en el diseño del sitio en los clientes habituales. Tras algunos años, los clientes registrados desarrollan determinados hábitos con respecto a la organización del sitio. Los comentarios de los usuarios registrados no son tan positivos como los de los usuarios no registrados y a algunos de ellos no les gustan los cambios en absoluto. El negocio necesita conocer este problema y considerar si un cambio atraerá a la cantidad de nuevos clientes suficiente como para arriesgarse a perder algunos de los clientes actuales.