UEA Datos a Gran Escala

Enfoque Metodológico Basado en la Minería de Datos (*Data Mining*)

Dr. Pedro Pablo González Pérez

e-mail: pgonzalez@correo.cua.uam.mx

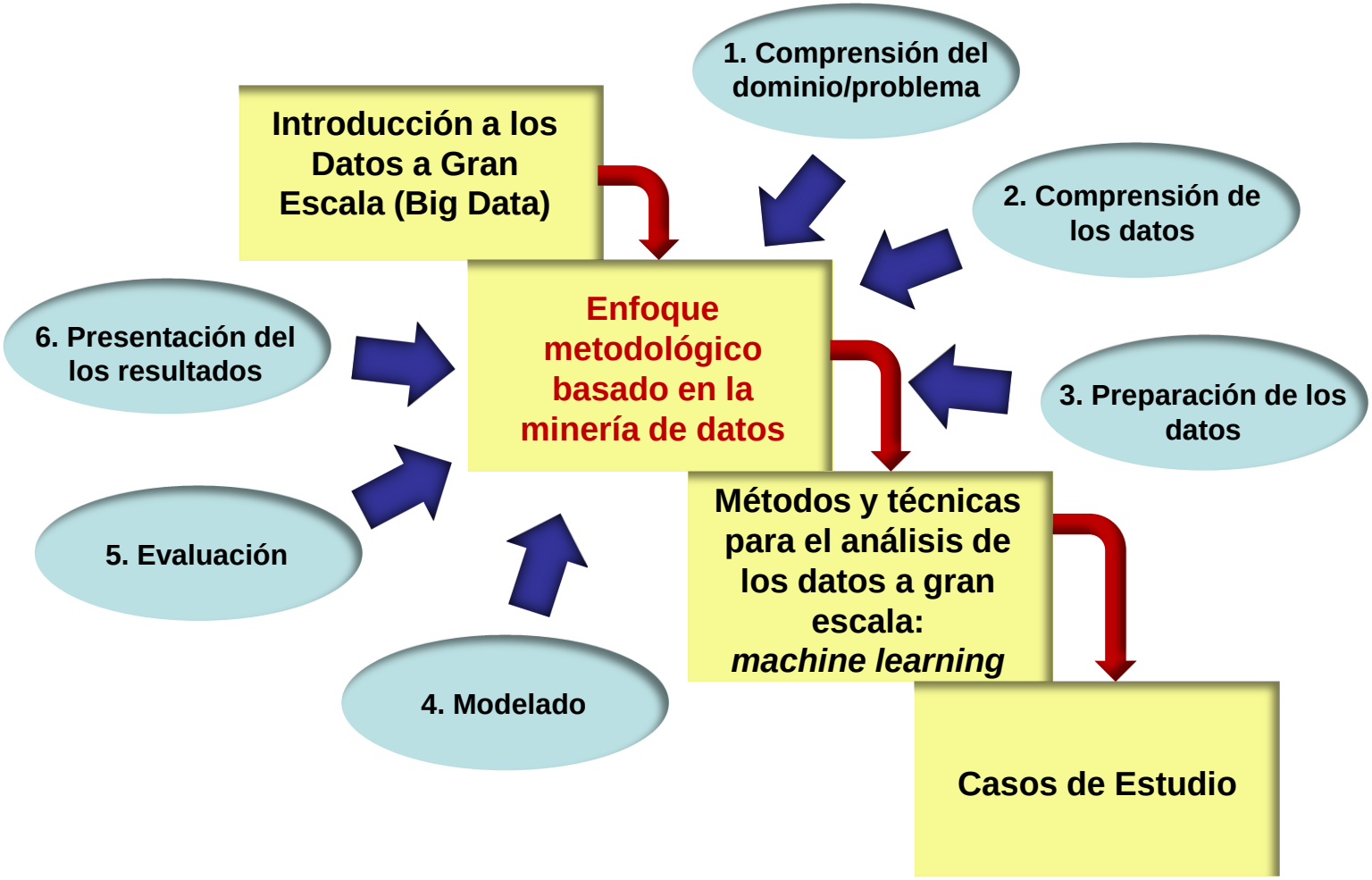
http://dcni.cua.uam.mx/division/usuario?p=31#

Departamento de Matemáticas Aplicadas y Sistemas

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA
Unidad Cuajimalpa

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos



Enfoque metodológico basado en la minería de datos

Preprocesamiento y Análisis de los Datos: Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000]

- *CRISP-DM* (*Cross Industry Standard Process for Data Mining*) es una metodología para minería de datos, propuesta en [Shearer, 2000]. A pesar de no constituir un enfoque metodológico reciente, éste es aun relevante como guía para proyectos de minería de datos, por lo que continúa siendo hoy en día ampliamente usada, difundida, extendida y adaptada a proyectos de datos a gran escala.

Enfoque metodológico basado en la minería de datos

Preprocesamiento y Análisis de los Datos: Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000; IBM Corporation, 2012]

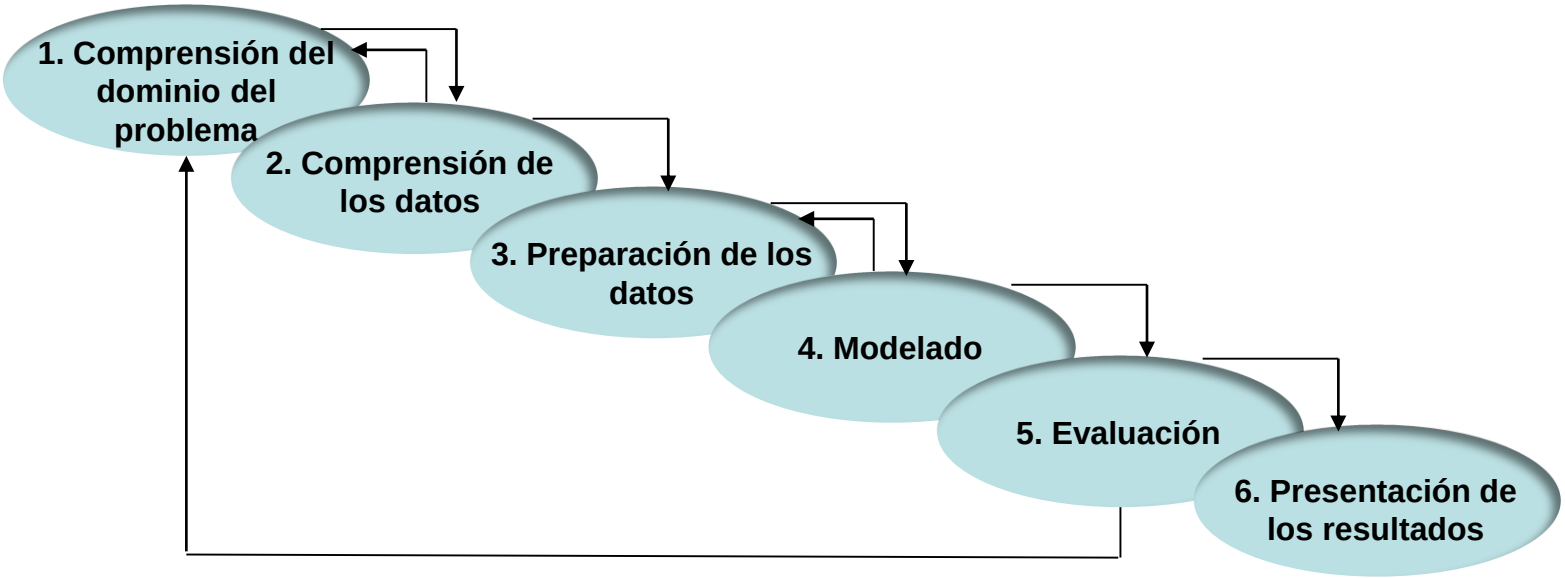
- CRISP-DM está compuesta por seis fases principales:
  1) **Comprensión del dominio del problema (negocio).**
  2) **Comprensión de los datos.**
  3) **Preparación de los datos.**
  4) **Modelado.**
  5) **Evaluación.**
  6) **Despliegue.**
- Como se puede apreciar en la siguiente figura, el modelo de fases de CRIPS-DM no es precisamente un ciclo de Cascada Pura, ya que la secuencia entre las fases no es estricta, siendo posible retroceder a fases anteriores, siempre que sea necesario.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

Preprocesamiento y Análisis de los Datos: Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000; IBM Corporation, 2012]



Enfoque metodológico basado en la minería de datos

Preprocesamiento y Análisis de los Datos: Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000]

- *CRISP-DM* es una metodología de minería de datos muy flexible, la cual se puede adaptar fácilmente a las necesidades concretas o al problema específico, relacionados con el volumen de datos a procesar y/o analizar.
- En dependencia de tales necesidades concretas o problema específico a resolver, en ocasiones las fases de **comprensión del problema** y **preparación de los datos** pudieran llegar a ser las más relevantes, mientras que en otras ocasiones serán las fases de **modelado** y **evaluación** las más significativas, para la toma de decisiones.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1. Comprensión del dominio del problema

1) Comprensión del dominio del problema o negocio.

La fase "**Comprensión del dominio del problema o negocio**" resulta de gran importancia y ejerce un gran impacto en las fases sucesivas, independientemente de la metodología o enfoque de minería de datos que se haya seleccionado. Esto se debe a que es en esta fase donde se define de forma clara el problema que se intenta resolver, se focaliza en la comprensión de las metas/objetivos del trabajo a desarrollar y proporciona una perspectiva de minería de datos para comprender cuáles datos deben ser analizados.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1) Comprensión del dominio del problema o negocio.

Comúnmente, la comprensión del dominio del problema o negocio puede ser llevada a cabo ejecutando las siguientes actividades:

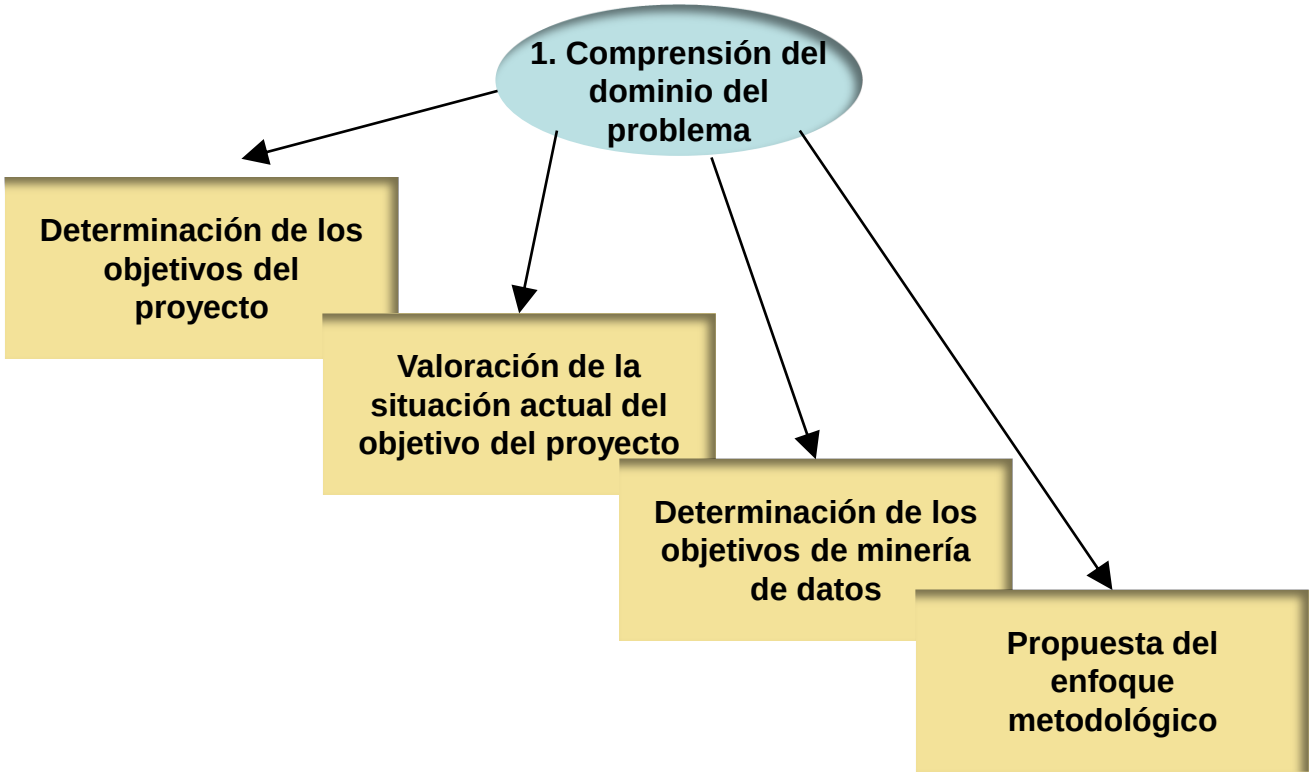
- **Determinación de los objetivos del proyecto.**
- **Valoración de la situación actual del objetivo del proyecto.**
- **Determinación de los objetivos de minería de datos.**
- **Propuesta del enfoque metodológico para desarrollar el proyecto.**

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1) Comprensión del dominio del problema o negocio.



Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1) Comprensión del dominio del problema o negocio.

- **Determinación de los objetivos del proyecto.**
  - En la fase "**Comprensión del dominio del problema**" es imprescindible conocer de forma clara las razones que justifican llevar a cabo el proyecto o estudio. De aquí que, un aspecto clave de esta fase sea precisamente la determinación de los objetivos del proyecto.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1. Comprensión del dominio del problema

1) Comprensión del dominio del problema o negocio.

- **Determinación de los objetivos del proyecto.**
  - Los objetivos del proyecto dependerán completamente del dominio del problema o negocio en el cual se enmarca el proyecto. Ejemplos genéricos de dominio del problema o negocio son los siguientes:
    - **Ventas y mercadotecnia (*marketing*).**
    - **Finanzas y mercado bursátil.**
    - **Otorgamiento de créditos.**
    - **Lotería y juegos de apuesta.**
    - **Clima y contaminación atmosférica.**
    - **Investigación científica.**
    - **Otras ramas de las ciencias e ingenierías.**

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1. Comprensión del dominio del problema

1) Comprensión del dominio del problema o negocio.

- **Determinación de los objetivos del proyecto.**
  - Los objetivos del proyecto deben indicar claramente:
    - ¿Qué se pretende hacer?
    - ¿Cuál es el problema que se necesita resolver?
    - ¿Cuál es la pregunta que se requiere responder?

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1. Comprensión del dominio del problema

1) Comprensión del dominio del problema o negocio.

- **Determinación de los objetivos del proyecto.**
  - Ejemplos de objetivos de proyecto:
    - En el dominio de Ventas y Mercadotecnia, el objetivo podría ser **la Predicción** del Incremento de las Ventas y la Retención de Clientes, a partir de acciones tales como ofertas y descuentos.
    - En el dominio de Otorgamiento de Créditos, el objetivo podría ser **la Clasificación** de los Solicitantes de Crédito Bancario en Aceptados o Rechazados, según su historial crediticio, capacidad de pago, etc.
    - En el dominio de Clima y Contaminación Atmosférica, el objetivo podría ser **la Predicción** de la Calidad del Aire en determinadas zonas metropolitanas.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1. Comprensión del dominio del problema

1) Comprensión del dominio del problema o negocio.

- **Valoración de la situación actual del objetivo del proyecto.**
  - Una vez que se ha definido el objetivo del proyecto, entonces es necesario valorar la situación actual del mismo, considerando los siguientes aspectos:
    - ¿Se comprende de forma clara el problema que se intenta abordar?
    - ¿Existen datos disponibles para efectuar el análisis?
    - De contar con datos disponibles, ¿cuál es la fuente de estos datos y de qué tipo son?
    - ¿Se dispone de los recursos humanos y tecnológicos para desarrollar el proyecto?
    - ¿Se han identificado factores de riesgo que afecten el desarrollo del proyecto?

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1) Comprensión del dominio del problema o negocio.

- **Determinación de los objetivos de minería de datos.**
  - En esta fase es imprescindible conocer de forma clara las razones que justifican llevar a cabo el proyecto de minería de datos. De aquí que, un aspecto clave de la fase "**Comprensión del dominio del problema**" sea precisamente la determinación de los objetivos del proyecto y la traducción de éstos a objetivos de minería de datos.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1) Comprensión del dominio del problema o negocio.

- **Determinación de los objetivos de minería de datos.**
  - Los objetivos de minería de datos se refieren a qué tipo de información deseamos explorar y encontrar en los datos. Por ejemplo:
    - Los datos que satisfagan ciertas condiciones.
    - Relaciones existentes entre ciertas características o campos de los datos.
    - Clasificación o agrupamiento de los datos.
    - Predicción a partir de los propios datos.
    - Patrones.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1. Comprensión del dominio del problema

1) Comprensión del dominio del problema o negocio.

- **Determinación de los objetivos de minería de datos.**
  - Ejemplos de objetivos de minería de datos son los siguientes:
    - **Selección de los datos que satisfagan determinadas características.**
    - **Búsqueda de relaciones/correlaciones.**
    - **Clasificación/Agrupamiento (*clustering*).**
    - **Predicción/pronóstico.**
    - **Búsqueda de patrones.**
    - **…**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1) Comprensión del dominio del problema o negocio.

- **Propuesta del enfoque metodológico (plan de proyecto de minería de datos).**
  - El enfoque metodológico o plan de proyecto de minería de datos se propone a partir de todos los aspectos que se han identificado hasta el momento como parte de la fase "**Comprensión del dominio del problema o negocio**". Dentro de estos aspectos se encuentran **el objetivo del proyecto**, **la valoración de la situación actual del objetivo del proyecto**, y **los objetivos de minería de datos**.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1. Comprensión del dominio del problema

1) Comprensión del dominio del problema o negocio.

- **Propuesta del enfoque metodológico (plan de proyecto de minería de datos).**
  - El enfoque metodológico o plan de proyecto de minería de datos puede ser visto como un cronograma de trabajo (comúnmente en forma de tabla), el cual debe especificar, entre otros aspectos:
    - **Nombre de la fase de la metodología CRISP-DM.**
    - **Tiempo en semanas dedicado a dicha fase.**
    - **Recursos humanos y tecnológicos que requiere.**
    - **Posibles riesgos a mitigar en cada fase.**

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1. Comprensión del dominio del problema

1) Comprensión del dominio del problema o negocio.

- **Propuesta del enfoque metodológico (plan de proyecto de minería de datos) en forma de tabla.**

Fase	Tiempo a dedicar	Recursos humanos y tecnológicos	Riesgos atribuibles
Comprensión del dominio del problema			
Comprensión de los datos			
Preparación de los datos			
Modelado			
Evaluación			
Presentación			

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1) Comprensión del dominio del problema o negocio.

- Para el desarrollo de la fase "**Comprensión del dominio del problema**" no se requiere del apoyo de ninguna herramienta, aplicación o paquete de cómputo. Es suficiente elaborar un documento donde se registre la respuesta, resultado, descripción, argumentación, etc., a cada una de las actividades que engloba esta fase del enfoque de minería de datos **CRIPS-DM**.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

1) Comprensión del dominio del problema o negocio.

Estudiar, analizar y comprender cómo fue aplicada la fase "**Comprensión del dominio del problema**" en los siguientes casos de estudio, los cuales se encuentran disponibles como parte del material del curso:

- **Análisis financiero empresarial.**
- **Predicción del incremento de ventas en *e-commerce*.**
- **Predicción de ventas de inmuebles.**
- **Otorgamiento de crédito bancario.**
- **Diagnóstico COVID-19**

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

La fase "**Comprensión de los datos**" consiste en la descripción, exploración y análisis inicial de los datos, con la finalidad de:

- **Identificar si hay problemas de calidad (omisión, incompletitud, redundancia, falta de veracidad, etc.) presentes en los mismos.**
- **Descubrir y/o proponer ideas iniciales acerca de los datos.**
- **Establecer hipótesis acerca de la información que describen dichos datos.**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2) Comprensión de los datos.

- La fase **Comprensión de los datos** implica estudiar con mayor detenimiento los datos, tratando de comprender las posibles relaciones que se pudieran manifestar entre los mismos.
- La comprensión de los datos significa la exploración de los mismos, con el apoyo de **tablas**, **gráficos**, **resúmenes** y **otras herramientas estadísticas** que proporcionen diferentes vistas de los mismos.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

La fase **Comprensión de los datos** incluye las siguientes tareas:

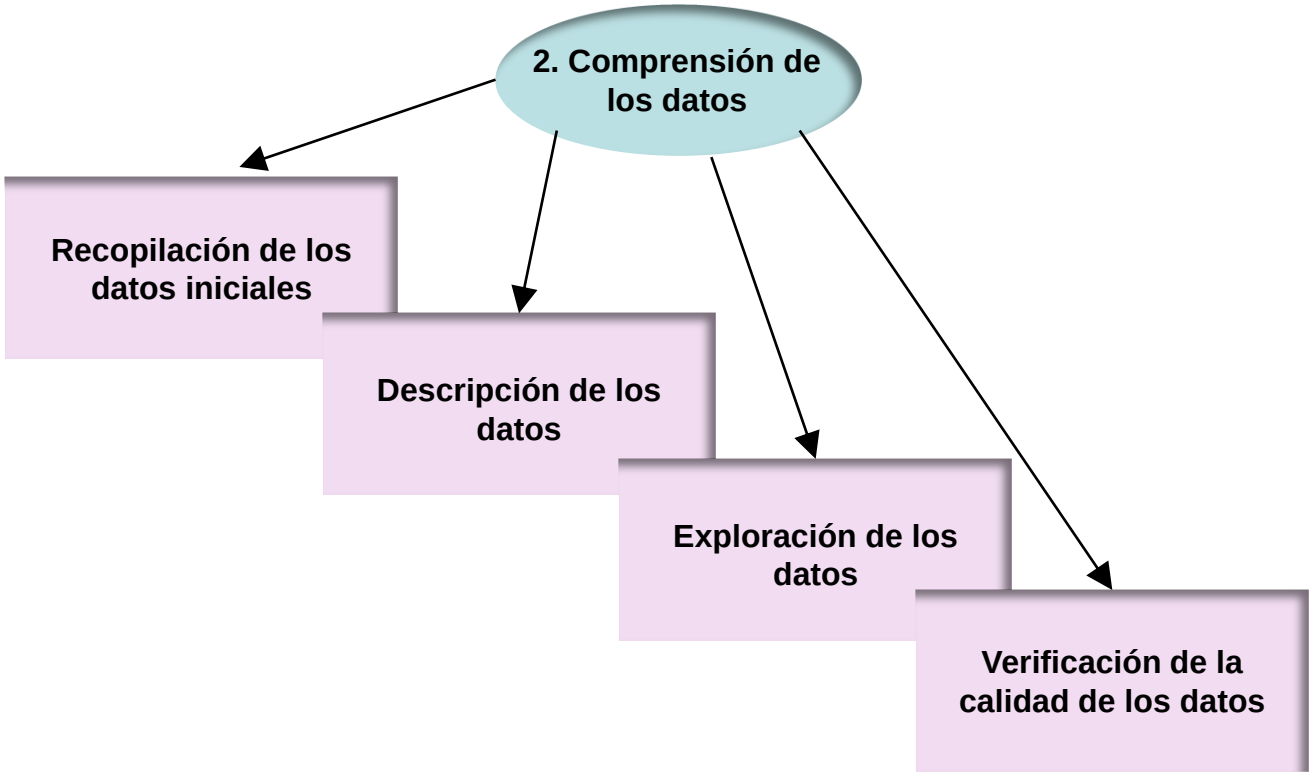
- **Recopilación de los datos iniciales.**
- **Descripción de los datos.**
- **Exploración de los datos.**
- **Verificación de la calidad de los datos.**

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2) Comprensión de los datos.



Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

- **Recopilación de los datos iniciales.**
  - Identificar las fuentes de las cuales provienen los datos:
    - **Datos existentes**. Son los datos con los que se cuenta inicialmente para el proyecto de minería de datos. Comúnmente, estos datos son obtenidos por la propia compañía, empresa, institución, grupo de investigación, etc.
    - **Datos adquiridos**. Se refiere a datos adquiridos a un tercero, y que pueden complementar los datos existentes, si se cuenta con ellos, o constituir los datos iniciales para el proyecto de minería de datos.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

- **Recopilación de los datos iniciales.**
  - Identificar las fuentes de las cuales provienen los datos:
    - **Datos adicionales**. De ser el caso, se deberá recurrir a otros datos que complementen las fuentes de datos ya mencionadas.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

- **Recopilación de los datos iniciales.**
  - Efectuar un análisis preliminar de los datos, centrando la atención en sus atributos (columnas):
    - ¿Cuáles son los atributos más prometedores?
    - ¿Cuáles son los atributos menos relevantes?, ¿Podrían excluirse del conjunto de datos?
    - ¿Se cuenta con datos suficientes?
    - ¿Son suficientes los atributos?

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

- **Descripción de los datos.**
  - Describir los datos con los que se cuenta, considerando los siguientes aspectos:
    - Cantidad de datos.
    - Cantidad de atributos o características.
    - Tipos de valores (numéricos, caracteres, booleanos). Identificar los tipos de datos simbólicos (fecha, hora, acceso a páginas Web, etc.) y proponer su conversión a datos numéricos.
    - De ser el caso, información sobre posibles clases o categorías que agrupan diferentes subconjuntos de estos datos, y cantidad de datos por clase.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

- **Exploración de los datos.**
  - La exploración de los datos constituye un primer análisis efectuado sobre el conjunto de datos para:
    - Comprender aún más dichos datos.
    - Corroborar o hacer aun más explícitos los objetivos de minería de datos.
    - Formular hipótesis sobre los datos.
    - Identificar tareas requeridas en la fase **Preparación de los datos**, tales como limpieza, eliminación de datos redundantes, completamiento de datos faltantes, etc.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

- **Exploración de los datos.**
  - La exploración de los datos se apoya en diferentes tipos de **gráficos**, **tablas**, **herramientas de visualización** y **resúmenes estadísticos** que proporcionan una mejor comprensión de los datos y, como resultado, identificar atributos claves y eliminar atributos irrelevantes.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

- **Verificación de la calidad de los datos.**
  - La verificación de la calidad de los datos se refiere a identificar errores, inconsistencias o incompletitud en el conjunto de datos. Esta actividad se focaliza en la identificación de los siguientes tipos de problemas:
    - Datos perdidos o vacíos.
    - Errores en los datos.
    - Errores de medición.
    - Errores de codificación.
    - Atributos redundantes o de escasa utilidad.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

- **Verificación de la calidad de los datos.**
  - Todos los problemas identificados durante la **Verificación de los datos**, deberán ser solucionados en la próxima fase **Preparación de los datos**, antes de que éstos sean proporcionados como insumo al modelo que permitirá su análisis final.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

- El desarrollo de la fase "**Comprensión de los datos**" puede ser soportado en herramientas, aplicaciones y paquetes de cómputo tales como:
  - **Excel.**
  - **PSPP – Software libre.**
  - **IBM SPSS – Software de licencia.**
  - **IBM SPSS Modeler – Software de licencia (se puede utilizar la versión de prueba por 30 días).**
  - **R Studio – Software libre.**
  - **Cualquier otro paquete estadístico que ofrezca una amplia gama de gráficos y estadísticas.**

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

2. Comprensión de los datos

2) Comprensión de los datos.

Estudiar, analizar y comprender cómo fue aplicada la fase "**Comprensión de los datos**" en los siguientes casos de estudio, los cuales se encuentran disponibles como parte del material del curso:

- **Análisis financiero empresarial.**
- **Predicción del incremento de ventas en *e-commerce*.**
- **Predicción de ventas de inmuebles.**
- **Otorgamiento de crédito bancario.**
- **Diagnóstico COVID-19**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3) Preparación de los datos.

- Como su nombre lo indica, la fase "**Preparación de los datos**" prepara o da forma a los datos sin procesar para construir el conjunto final de datos que servirá como insumo para la construcción del modelo.
- De existir problemas de calidad presentes en los datos, tales como omisión, errores en los datos, errores de medición, errores de codificación, incompletitud, redundancia, falta de veracidad, etc., es precisamente en esta fase donde deben ser solucionados.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3) Preparación de los datos.

- La fase "**Preparación de los datos**" es comúnmente la fase que más tiempo consume, no sólo en la metodología CRISP-DM, sino en cualquier enfoque de minería de datos.
- Es en esta fase donde los "datos en bruto" deben ser preprocesados y convertidos en un conjunto de datos sobre el cual se puedan aplicar los objetivos de la minería de datos.

Enfoque metodológico basado en la minería de datos

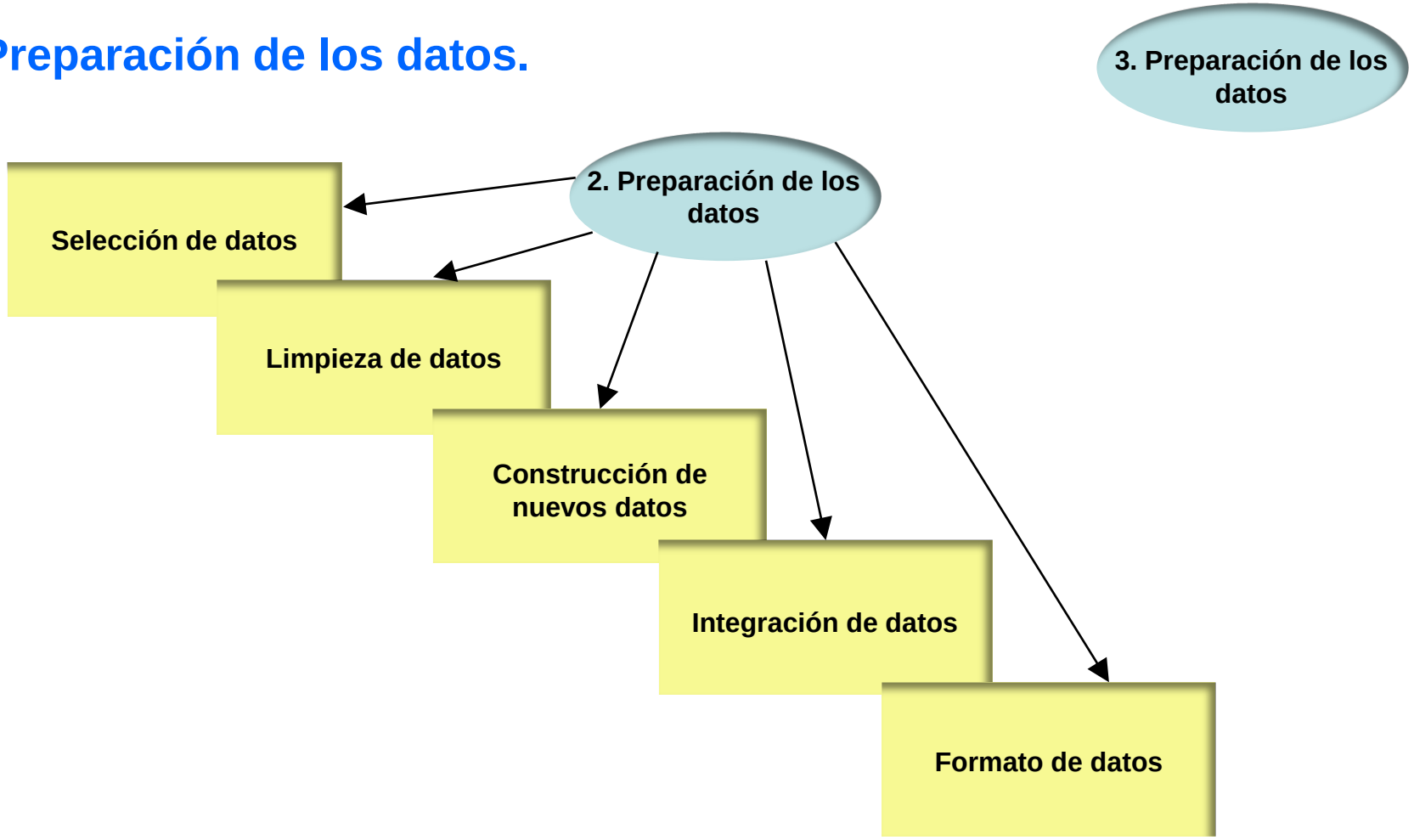
CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3) Preparación de los datos.

- La fase "**Preparación de los datos**" comúnmente requiere de las siguientes tareas:
  - **Selección de datos.**
  - **Limpieza de datos.**
  - **Construcción de nuevos datos.**
  - **Integración de datos.**
  - **Formato de datos.**

CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000; IBM Corporation, 2012]

3) Preparación de los datos.



Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3. Preparación de los datos

3) Preparación de los datos.

- **Selección de datos.**
  - Una vez llevada a cabo la recolección de datos iniciales (fase "**Comprensión de los datos**"), es posible iniciar la selección de los datos que satisfagan los objetivos de minería de datos propuestos. La selección de datos abarca dos criterios fundamentales:
    - **Selección de registros (filas)**. Se selecciona el subconjunto o subconjuntos de datos a utilizar. Comúnmente, no se requiere de todo el volumen de datos para el análisis y toma de decisiones. Mientras mayor sea el volumen de datos, mayor será el tiempo requerido para las fases de "**Preparación de los datos**" y "**Modelado**".

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3. Preparación de los datos

3) Preparación de los datos.

- **Selección de datos.**
  - **Selección de atributos o características (columnas)**. Se seleccionan los atributos o características más relevantes para el análisis y toma de decisiones. Comúnmente, no todos los atributos o características resultarán relevantes para el posterior análisis.

- La gran mayoría de las herramientas que soportan parcial o completamente la minería de datos, ofrecen las opciones de "**selección de datos**", tanto para la selección de registros (filas) como de atributos o características (columnas).

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3. Preparación de los datos

3) Preparación de los datos.

- **Limpieza de datos.**
  - La limpieza de datos se refiere a la corrección de problemas detectados en los datos (comúnmente, durante la fase "**Comprensión de los datos**") tales como:
    - **Datos perdidos (datos nulos o datos vacíos).**
    - **Errores en los datos.**
    - **Errores en la codificación de los datos.**
    - **Errores en la unidad de medida de los datos.**
    - **Errores en los metadatos.**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3) Preparación de los datos.

- **Limpieza de datos.**
  - En esta actividad se requiere identificar qué tipo de ruido contiene los datos y qué métodos utilizar para eliminar el ruido.
  - La gran mayoría de las herramientas que soportan parcial o completamente la minería de datos, ofrecen las opciones de "**limpieza de datos**", siendo una de las opciones más valiosas, el **reemplazo de datos perdidos**, tomando para ello el valor media, moda o mediana del atributo, según su tipo, continuo, nominal u ordinal, respectivamente.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3. Preparación de los datos

3) Preparación de los datos.

- **Construcción de nuevos datos.**
  - La construcción de nuevos datos se refiere a incorporar nuevos registros (filas) o nuevos atributos o características (columnas) como parte del conjunto de datos existentes.
    - La **incorporación de nuevos registros** (**filas**) toma lugar cuando el volumen de datos actual es insuficiente para alcanzar los objetivos de minería de datos.
    - La **incorporación de nuevos atributos o características** (**columnas**) toma lugar cuando en el volumen de datos actual faltan atributos relevantes para alcanzar los objetivos de minería de datos.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3) Preparación de los datos.

- **Construcción de nuevos datos.**
  - Comúnmente, la incorporación de nuevos atributos se traduce en la derivación de nuevos atributos a partir de los atributos existentes.
  - La gran mayoría de las herramientas que soportan parcial o completamente la minería de datos, ofrecen opciones de "**construcción de nuevos datos**", tanto para la creación de nuevos registros como de nuevos atributos o características.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3) Preparación de los datos.

- **Integración de datos.**
  - La integración de datos se refiere al incremento del conjunto de datos, a partir de datos provenientes de otras fuentes (por ejemplo, otras tablas pertenecientes a la misma base de datos, tablas procedentes de otras fuentes, etc.). Esta integración de datos puede darse ya sea incorporando nuevos atributos (columnas) o nuevos registros (filas), siendo estos dos enfoques la base de los dos métodos conocidos para la integración de datos: la fusión y la adición.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3) Preparación de los datos.

- **Integración de datos.**
  - **Fusión de datos**. Se refiere a la unión de **dos conjuntos de datos con registros similares** y atributos diferentes. Para fusionar los datos, se requiere que los dos registros a unir posean el mismo identificador llave. La fusión de datos resulta en un incremento de los atributos (o columnas) del conjunto de datos.
  - **Adición de datos**. Se refiere a la unión de **dos conjuntos de datos con atributos similares** y registros diferentes. La adición de datos resulta en un incremento de los registros (o filas) del conjunto de datos.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3. Preparación de los datos

3) Preparación de los datos.

- **Formato de datos.**
  - El formato de datos se refiere a considerar si el modelo que posteriormente se aplicará en la fase "**Modelado**" requiere de algún orden o clasificación particular ya sea de los registros o de los atributos.
  - De requerirse algún formato particular del conjunto de datos, la gran mayoría de las herramientas para el procesamiento de datos lo permiten.
  - Por ejemplo para un modelo de clasificación supervisada, sería conveniente que los atributos representando "clases" correspondan a las últimas columnas del conjunto de datos.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3) Preparación de los datos.

- **Formato de datos.**
  - El formato final de los datos dependerá estrechamente del modelo de análisis a utilizar. Problemas de predicción, problemas de clasificación y problemas de *clustering*, podrían requerir formatos de datos diferentes.
  - Sin embargo, en muchas ocasiones, si se ha efectuado un arduo trabajo en esta fase "**Preparación de los datos**", pensando en el tipo de modelo de análisis, entonces el formato de los datos ya ha sido concebido de forma implícita.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3) Preparación de los datos.

- El desarrollo de la fase "**Preparación de los datos**" puede ser soportado en herramientas, aplicaciones y paquetes de cómputo tales como:
  - **Excel.**
  - **PSPP – Software libre.**
  - **IBM SPSS – Software de licencia.**
  - **IBM SPSS Modeler – Software de licencia (se puede utilizar la versión de prueba por 30 días).**
  - **R Studio – Software libre.**
  - **Cualquier otro paquete estadístico que ofrezca una amplia gama de gráficos y estadísticas.**

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

3. Preparación de los datos

3) Preparación de los datos.

Estudiar, analizar y comprender cómo fue aplicada la fase "**Preparación de los datos**" en los siguientes casos de estudio, los cuales se encuentran disponibles como parte del material del curso:

- **Análisis financiero empresarial.**
- **Predicción del incremento de ventas en *e-commerce*.**
- **Predicción de ventas de inmuebles.**
- **Otorgamiento de crédito bancario.**
- **Diagnóstico COVID-19**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

- La fase "**Modelado**" no debería ser iniciada, si alguna de las siguientes cuestiones no ha sido manejada o considerada de forma previa:
  - ¿Se ha garantizado un fácil acceso a los datos desde las herramientas de modelado a utilizar?
  - ¿A través de la comprensión de los datos y de la exploración de éstos se ha podido identificar el subconjunto de datos más prometedores?
  - ¿Se ha efectuado una adecuada limpieza de los datos?

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

- ¿Se conocen las herramientas de modelado?
- ¿Los datos se encuentran en el formato requerido por el modelo?
- En caso de haber efectuado integración de datos, ¿existe algún problema en la unión y/o fusión de los datos?

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

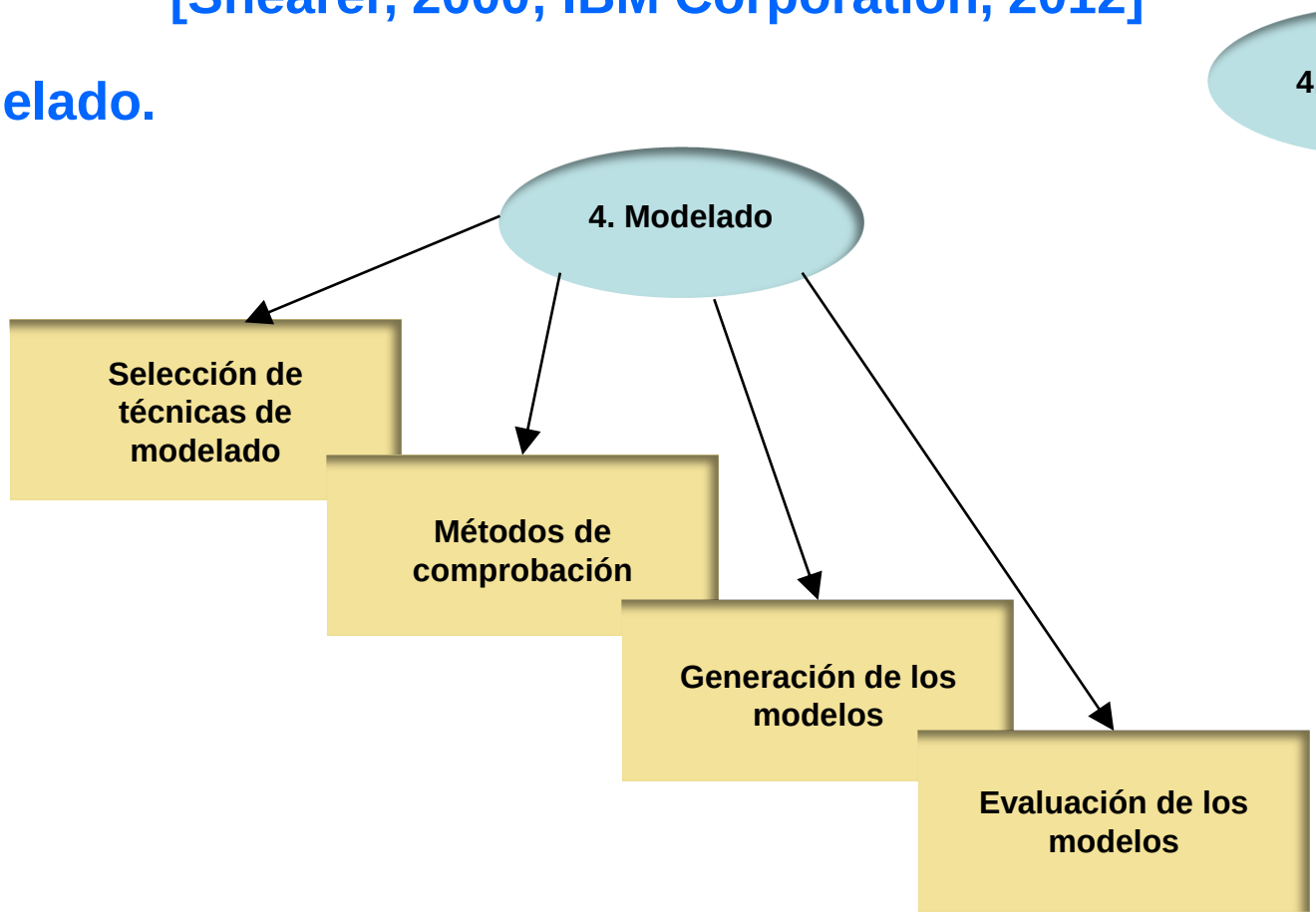
4) Modelado.

- Comúnmente, cada uno de los modelos seleccionados se ejecuta inicialmente con los propios parámetros "por defecto" propuestos por el modelo. Posteriormente, estos parámetros pueden ser ajustados para ver si es posible mejorar los resultados (bondad, error, coeficiente de correlación etc.) que produce el modelo.
- La fase "**Modelado**" incluye las siguientes actividades:
  - Selección de técnicas de modelado.
  - Métodos de comprobación.
  - Generación de los modelos.
  - Evaluación de los modelos.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000; IBM Corporation, 2012]

4) Modelado.



4. Modelado

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

- **Selección de técnicas de modelado**
  - La fase "**Modelado**" comúnmente se ejecuta utilizando varios modelos, de forma tal que, para el conjunto de datos, se puedan obtener resultados producidos por varios métodos o técnicas, y, de esta forma, poder comparar cuál fue el modelo que produjo los mejores resultados para el objetivo de minería de datos.
  - La elección de los modelos a utilizar depende estrechamente del objetivo de minería de datos propuesto. Por ejemplo:
    - Si se trata de un problema de predicción, entonces será necesario utilizar técnicas o métodos basado en la regresión o aprendizaje supervisado.
    - Si se trata de un problema de clasificación, entonces lo adecuado sería utilizar algún modelo de aprendizaje supervisado, tal como el *perceptron* multicapas o algún otro modelo similar.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

- **Selección de técnicas de modelado**
  - Para la selección del modelo (o de los modelos) más adecuado se deben considerar los siguientes aspectos:
    - Los tipos de datos con los que se cuenta para la minería de datos. Es decir, considerar si son datos continuos, categóricos, ordinales, nominales, etc. Es muy común que cada tipo de modelo imponga restricciones sobre el tipo de dados, tanto sobre los datos-predictores como sobre los datos objetivo.
    - Los objetivos de minería de datos. Por ejemplo, problema de predicción, problema de clasificación, problema de agrupamiento, etc.
    - Requerimientos específicos del modelado. Por ejemplo, ¿qué tipos de resultados necesita que proporcione el modelo y cómo requiere que se presenten?

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4. Modelado

4) Modelado.

- **Selección de técnicas de modelado (I)**
  - Entre las técnicas comúnmente utilizadas para construir el modelo que analizará los datos se encuentran:
    - **Árboles de decisión supervisados, para problemas de predicción o clasificación.**
    - **Modelos de regresión, para problemas de predicción: regresión lineal, regresión logística, etc.**
    - **Redes neuronales supervisadas, para problemas de predicción o clasificación: *Perceptron* Multi-Capa, Red neuronal SVM (*Support Vector Machine*), Red neuronal LSVM (*Linear Support Vector Machine*), Máquina de Boltzmann Restringida (*Restricted Boltzmann Machine*).**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

- **Selección de técnicas de modelado (II)**
  - **Redes neuronales no supervisadas, para problemas de agrupamiento (*clustering*): Red neuronal de Kohonen.**
  - **Algoritmo K-NN (*K-Nearest-Neighbor*), para problemas de predicción o clasificación.**
  - **Aprendizaje profundo (*deep learning*) para problemas de predicción o clasificación.**

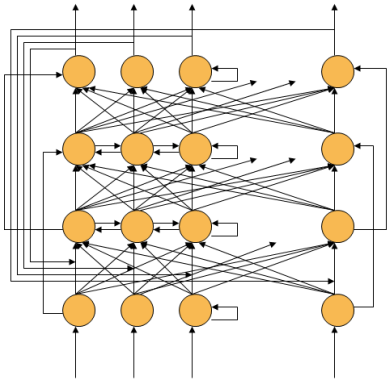
CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

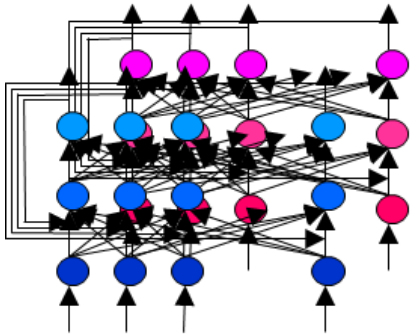
- **Selección de técnicas de modelado**

Datos a Gran Escala (Big Data)

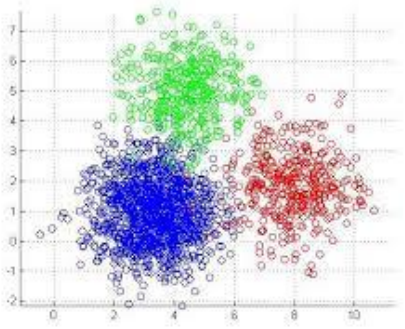
Modelos de Redes Neuronales Artificiales



Modelos de Aprendizaje Profundo (*Deep Learning*)

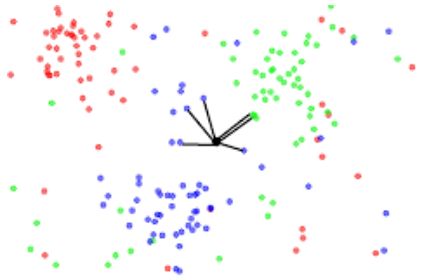


Algoritmos de Agrupamiento



Métodos estadísticos

Algoritmos KNN



Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

- **Métodos de comprobación**
  - Los métodos de comprobación se refieren a la forma en que se comprobarán los resultados producidos por el modelo.
  - Un método de comprobación de un modelo debe incluir:
    - **El criterio de bondad del modelo**. La bondad del modelo se refiere a una medición del desempeño del modelo. Por ejemplo, para modelos supervisados, el criterio de bondad comúnmente se refiere a la tasa de error del modelo; mientras que para modelos no supervisados, la bondad puede referirse a facilidad de interpretación, tiempo de procesamiento, entre otros aspectos.
    - **Definición de los datos para comprobar el criterio de bondad.**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4. Modelado

4) Modelado.

- **Métodos de comprobación**
  - **Definición de los datos para comprobar el criterio de bondad.** Para comprobar el criterio de bondad del modelo se requiere contar con nuevos datos. Es decir, datos que no hayan sido suministrados al modelo como insumo, y para los cuales comúnmente se conoce o sospecha el resultado esperado, en términos de clase, *cluster* o resultado predictivo.
- Los métodos de comprobación permiten constatar los resultados producidos por cada uno de los modelos seleccionados, antes de decidir cuál o cuáles de éstos se utilizarán.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

- **Métodos de comprobación**
  - En la actualidad, muchas herramientas de minería de datos incluyen entre sus prestaciones la partición de los datos en dos conjuntos, uno para el entrenamiento del modelo y el otro para la prueba o verificación del mismo.
  - Como ya se mencionó, todos los modelos supervisados (tales como redes neuronales con aprendizaje supervisado, árboles de decisión supervisados, algoritmos de predicción supervisados, entre otros) requieren necesariamente de un conjunto de datos de prueba para comprobar el desempeño del modelo.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4. Modelado

4) Modelado.

- **Generación de los modelos**
  - La **generación de los modelos** se refiere a la selección final y ejecución de los modelos seleccionados para analizar los datos.
  - Afortunadamente, existe una extensa gama de herramientas para el análisis inteligente de datos y minería de datos que incluyen una amplia variedad de modelos de análisis – tales como como árboles de búsqueda, modelos de predicción, algoritmos de agrupamiento, redes neuronales artificiales, entre otras técnicas – implementados y disponibles para su uso.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

- **Generación de los modelos**
  - La **generación de un modelo** comúnmente incluye tareas tales como:
    - Selección y preparación de los datos para el modelo particular.
    - Configuración de los parámetros del modelo.
    - Ejecución del modelo.
    - Exploración de los resultados del modelo.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

- **Evaluación de los modelos**
  - Una vez seleccionados los modelos a utilizar y ejecutados los mismos sobre el conjunto de datos a analizar, la **evaluación de los modelos** se refiere a observar los resultados producidos por cada uno de éstos, para decidir cuál resultó ser el modelo más preciso – en el caso de modelos supervisados, aquel con el menor error producido.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

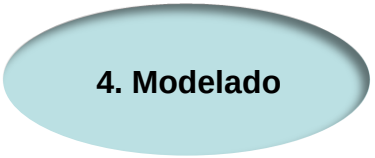
4) Modelado.

- **Evaluación de los modelos**
  - Un método comúnmente utilizado para la evaluación de los modelos es la exploración de los resultados producidos, a través de gráficos, tablas y métricas de desempeño.
  - Una vez evaluados los modelos, éstos pueden ser clasificados a partir de criterios tales como efectividad o precisión, tiempo de ejecución, facilidad para la interpretación de los resultados, etc.

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]



4) Modelado.

- El desarrollo de la fase "**Modelado**" puede ser soportado en herramientas, aplicaciones y paquetes de cómputo tales como:
  - **Excel. Sólo de forma parcial, para problemas de predicción.**
  - **PSPP – Software libre. Sólo de forma parcial, para problemas de predicción.**
  - **IBM SPSS – Software de licencia. Sólo de forma parcial, para problemas de predicción.**
  - **IBM SPSS Modeler – Software de licencia (se puede utilizar la versión de prueba por 30 días).**
  - **R Studio – Software libre.**
  - **Librerías dedicadas a la minería de datos con herramientas basadas en *machine learning* y en otras técnicas de inteligencia artificial.**

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

4) Modelado.

Estudiar, analizar y comprender cómo fue aplicada la fase "**Modelado**" en los siguientes casos de estudio, los cuales se encuentran disponibles como parte del material del curso:

- **Análisis financiero empresarial.**
- **Predicción del incremento de ventas en *e-commerce*.**
- **Predicción de ventas de inmuebles.**
- **Otorgamiento de crédito bancario.**
- **Diagnóstico COVID-19**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000; IBM Corporation, 2012]

5) Evaluación.

- En la fase "**Evaluación**" los resultados producidos por el modelo de análisis son evaluados, para asegurar que los objetivos del proyecto – por ejemplo, objetivos comerciales de la organización – sean alcanzados.
- La fase "**Evaluación**" se refiere a **la evaluación e interpretación de los resultados y descubrimientos producidos por el modelo**, y no a la evaluación del modelo en sí, ya que esta última actividad se llevó a cabo al final de la fase "**Modelado**".
- Para ejecutar la fase "**Evaluación**" se debe comprender claramente los objetivos comerciales de la organización o, en su caso, los objetivos del tipo de proyecto en cuestión.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

5) Evaluación.

- Por ejemplo, para los objetivos comerciales relacionados con el posicionamiento, rentabilidad y retención de clientes de una plataforma *e-commerce*, los resultados globales podrían ser:

  - **Recomendaciones sobre mejoras en los productos ofrecidos a los clientes.**
  - **Recomendaciones sobre mejoras en el mecanismo de ventas cruzadas, de forma tal que el cliente no perciba un comportamiento voraz por parte del negocio.**
  - **Recomendaciones sobre mejoras en el sitio Web, de forma tal que ofrezca una mejor navegación e interacción al usuario.**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

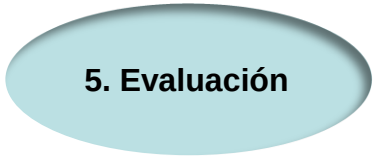
5) Evaluación.

- Por ejemplo, para los objetivos comerciales de un estudio de mercadotecnia, relacionado con el incremento de las ventas de bienes de consumo a partir de la ejecución de determinadas promociones, los resultados globales podrían ser:

  - **Recomendaciones de mejoras para que la promoción resulte mucho más atractiva, en aquellos tipos de bienes de consumo donde no se obtuvieron los resultados previstos.**
  - **Recomendaciones sobre cuáles variaciones efectuar en el monto de la promoción, en dependencia de los resultados obtenidos para cada tipo de bien de consumo.**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

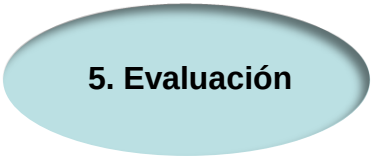


5) Evaluación.

- Los siguientes aspectos pueden proporcionar un gran soporte para la evaluación de los resultados producidos por el modelo:
  - **Claridad con la cuál se expresan los resultados del modelo.**
  - **Facilidad y claridad para la presentación de los resultados del modelo.**
  - **Descubrimientos relevantes efectuados a partir de los resultados producidos por el modelo.**
  - **Correspondencia de los resultados producidos por el modelo a los objetivos del proyecto (por ejemplo, objetivos comerciales de la compañía).**
  - **Otras conclusiones adicionales generadas en los resultados producidos por el modelo.**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]



5) Evaluación.

- El desarrollo de la fase "**Evaluación**" requiere de una mayor argumentación, discusión y documentación. Sin embargo, de ser necesario, se pueden utilizar las herramientas, aplicaciones y paquetes de cómputo relacionados en la fase "**Modelado**".

UEA Datos a Gran Escala

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

5) Evaluación.

Estudiar, analizar y comprender cómo fue aplicada la fase "**Evaluación**" en los siguientes casos de estudio, los cuales se encuentran disponibles como parte del material del curso:

- **Análisis financiero empresarial.**
- **Predicción del incremento de ventas en *e-commerce*.**
- **Predicción de ventas de inmuebles.**
- **Otorgamiento de crédito bancario.**
- **Diagnóstico COVID-19**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000; IBM Corporation, 2012]

6) Despliegue.

- La fase "**Despliegue**" se refiere a la presentación final de los resultados y a utilizar los nuevos conocimientos contenidos en éstos para responder al objetivo del proyecto – por ejemplo, en el caso de objetivos comerciales, podría ser efectuar mejoras en la organización o compañía, tales como un mejor posicionamiento, incremento de la rentabilidad, retención de clientes, etc.
- Es en la fase "**Despliegue**" donde se decide si los nuevos patrones descubiertos en los resultados de los modelos son lo suficientemente fuertes como para conllevar a grandes cambios en las estrategias de la organización, cuando se trata de objetivos comerciales.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Shearer, 2000; IBM Corporation, 2012]

6) Despliegue.

- La fase "**Despliegue**" significa:
  - Cómo utilizar los resultados producidos por la minería de datos para efectuar modificaciones y mejoras en la organización, cuando se trata de objetivos comerciales.
  - Cómo utilizar los resultados producidos por la minería de datos para guiar fases claves de un proyecto de investigación – por ejemplo, la evaluación experimental, tomando como guía la predicción efectuada por el modelo – cuando se trata de objetivos de investigación científica.

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

6. Despliegue

5) Despliegue.

- Por ejemplo, para los objetivos comerciales relacionados con el posicionamiento, rentabilidad y retención de clientes de una plataforma *e-commerce*, acciones a tomar durante la fase "**Despliegue**" podrían ser:

  - **Efectuar mejoras en los productos ofrecidos a los clientes.**
  - **Efectuar mejoras en el mecanismo de ventas cruzadas, de forma tal que el cliente no perciba un comportamiento voraz por parte del negocio.**
  - **Efectuar mejoras en el sitio Web, de forma tal que ofrezca una mejor navegación e interacción al usuario.**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

6. Despliegue

5) Despliegue.

- Por ejemplo, para los objetivos comerciales de un estudio de mercadotecnia, relacionado con el incremento de las ventas de bienes de consumo a partir de la ejecución de determinadas promociones, acciones a tomar durante la fase "**Despliegue**" podrían ser:
  - **Efectuar mejoras para que la promoción resulte mucho más atractiva, en aquellos tipos de bienes de consumo donde no se obtuvieron los resultados previstos.**
  - **Efectuar variaciones en el monto de la promoción, en dependencia de los resultados obtenidos para cada tipo de bien de consumo.**

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

6) Despliegue.

- **Creación del informe final**. Comúnmente, la fase "**Despliegue**" conlleva la elaboración de un **informe final** que incluya los siguientes aspectos:
  - Descripción detallada del problema original.
  - Resumen del procedimiento utilizado para llevar a cabo la minería de datos.
  - Resumen de los resultados del proyecto de minería de datos, incluyendo modelos, resultados, nuevos conocimientos, etc.
  - Resumen del plan propuesto para el "**despliegue**".
  - Recomendaciones para futuros proyectos de análisis inteligente de datos a gran escala.
  - ….

Enfoque metodológico basado en la minería de datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*)
[Shearer, 2000; IBM Corporation, 2012]

6) Despliegue.

- El desarrollo de la fase "**Despliegue**" requiere de una mayor argumentación, discusión y documentación. Sin embargo, de ser necesario, se pueden utilizar las herramientas, aplicaciones y paquetes de cómputo relacionados en la fase "**Modelado**".

UEA Datos a Gran Escala

Bibliografía necesaria o recomendable:

1. IBM. Manual CRISP-DM de IBM SPSS Modeler. IBM Corporation. 2012.
2. Joyanes Aguilar, L. Big data. Análisis de grandes volúmenes de datos en organizaciones. Alfaomega, 2013.
3. Marr, B. Big data in practice. How 45 successful companies used big data analytics to deliver extraordinary results. Wiley, 2016.
4. Marz, N., Warren, J. Big data: Principles and best practices of scalable realtime data systems. Manning Publications, 2015.
5. Mayer-Schönberger, V., Cukier, K. Big data. La revolución de los datos masivos. Turner Noema, 2013.
6. Mayer-Schönberger, V., Cukier, K. Big data: A revolution that will transform how we live, work, and think. John Murray, 2017.
7. Sinha, S. Making big data work for your business. Impackt Publishing, 2014.