Metodología CRISP y preparación de datos

Equipo Innovación Copec

Roberto Muñoz
Senior Data Scientist
Digital Data Analytics
EY

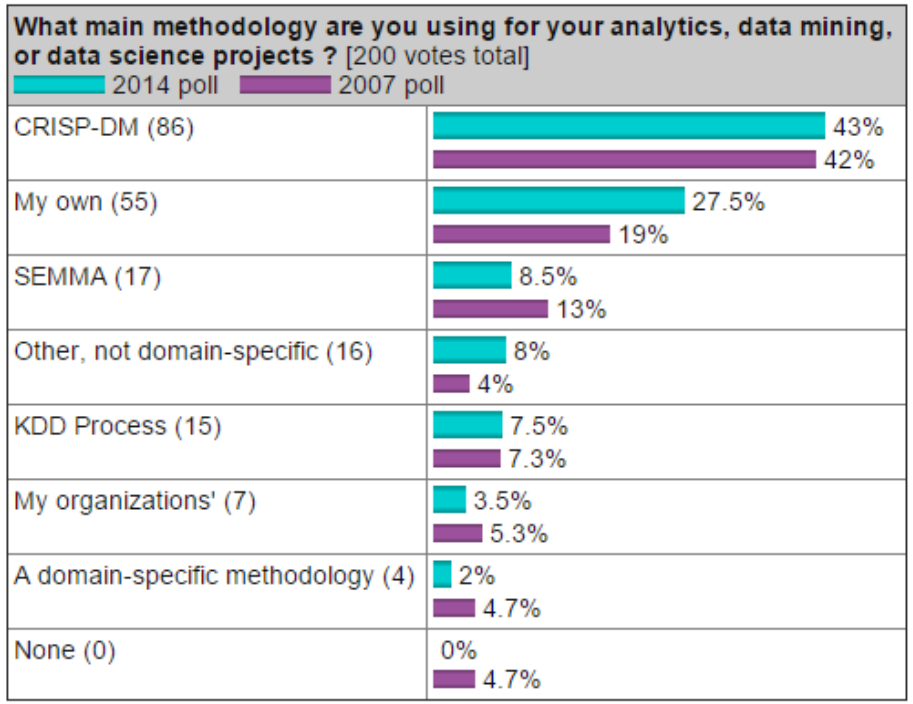
Una metodología

- Es un proceso preciso y formal.

- Una metodología incluye:
  - Actividades paso a paso para cada fase.
  - Roles individuales para cada actividad.
  - Productos y niveles de calidad para cada actividad.
  - Herramientas y técnicas que se usarán para cada actividad.

Metodologías más utilizadas para Análisis de Datos

- Distribución regional de los votantes
  - US/Canada 45.5%
  - Europe 28.5%
  - Asia 14.0%
  - Latin America 9.5%
  - Other 2.5%

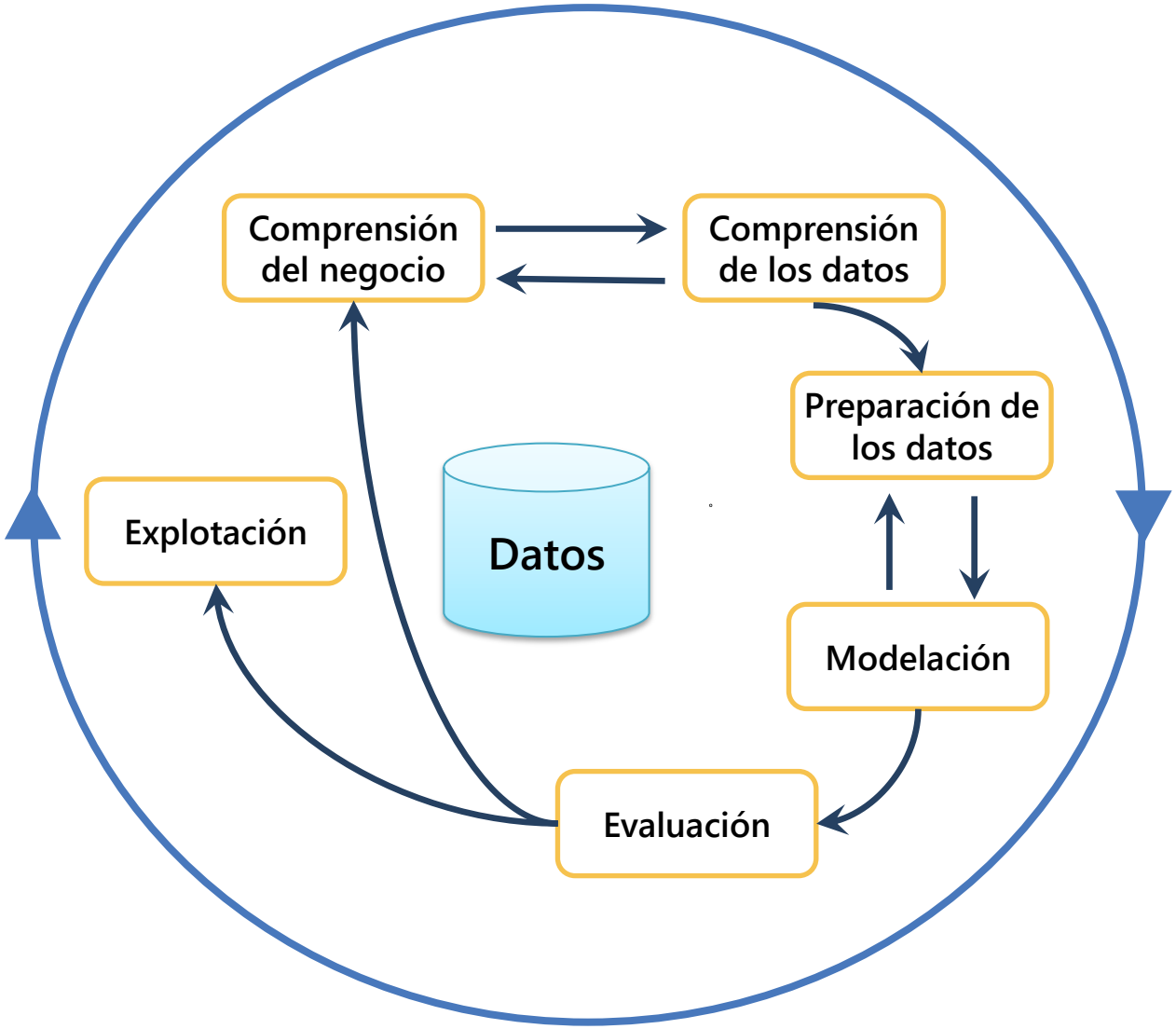


Fuente: KDnuggets Poll, Octubre 2014
http://www.kdnuggets.com

CRISP - DM

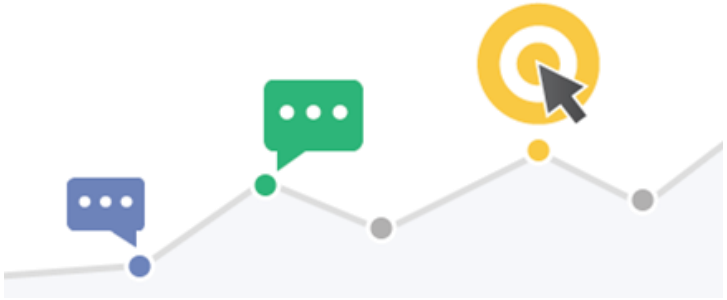
- CRoss-Industry Standard Process for Data Mining.
- Metodología para el proceso de Minería de Datos
  - Valida el proceso, ayuda a planear y administrar proyectos.
- Desarrollado el año 2000 por algunas compañías: SPSS/ISL, NCR, OHRA.
- Está enfocado en el negocio y al análisis técnico.

Visión General



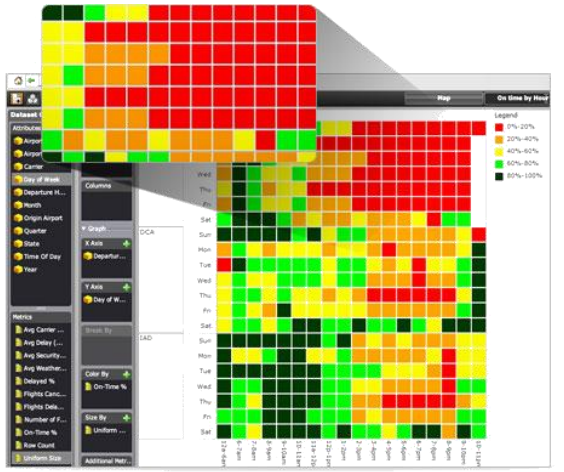
Fase 1: Comprensión del Negocio

- Determinar los objetivos de negocio
  - Dentro de este contexto es importante definir los criterios de éxito del negocio
- Levantamiento de requerimientos, riesgos, supuestos y beneficios
- Definir los objetivos del proyecto
  - Dentro de este contexto es importante definir los criterios de éxito del proyecto
- Generar planificación inicial



Fase 2: Comprensión de los Datos

- Objetivo:
  - Simplificar el problema y optimizar la eficiencia del modelo.



- ¿Cómo?
  - Uso de herramientas de visualización y técnicas de estadísticas descriptivas.

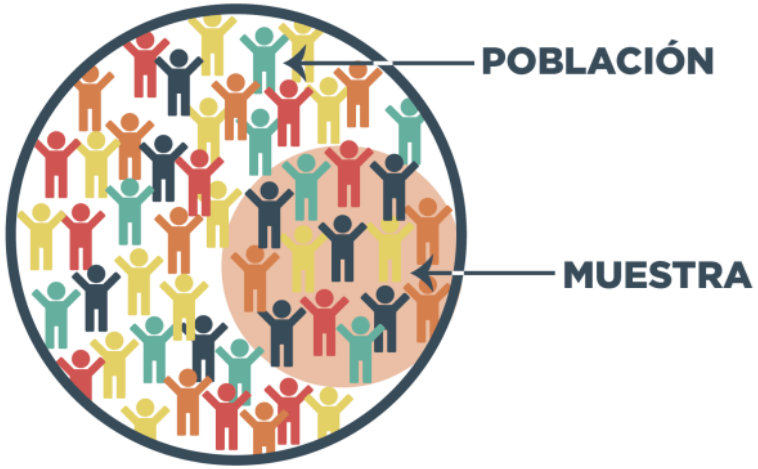
- Es relevante también determinar la calidad de los datos.

Fase 3: Preparación de los Datos

Selección

- Seleccionar el conjunto de datos o las variables o muestras sobre los cuales el proceso de análisis va a ser ejecutado.

- Selección de muestras.



Fase 3: Preparación de los Datos Limpieza de Datos

- La calidad del conocimiento a descubrir depende (además de otros factores) de la calidad de los datos analizados.

- Nuestro Objetivo:
  - Mejorar la calidad de los datos.

Fase 3: Preparación de los Datos

Limpieza: ¿En qué centrarse?

- Datos necesarios que no están a disposición
  - Estrategias para obtener datos
- Presencia de datos faltantes (missing values)
  - Estrategias para tratamiento de datos faltantes.
- Presencia de datos que no se ajustan al comportamiento general de los datos (outliers)

Fase 3: Preparación de los Datos

Missing values

- Es posible que los métodos que utilizaremos en fases posteriores no traten bien los campos con missing values.
- Hay que detectarlos y tratarlos.
- Posibles estrategias:
  - Ignorarlos
  - Eliminar variable
  - Filtrar registro
  - Reemplazar el valor
  - Etc.

Fase 3: Preparación de los Datos

Transformación de Datos

- Normalización de datos
- Construcción de nuevas variables que faciliten el proceso de minería de datos.
- Reducción de Dimensionalidad
  - Variables Correlacionadas
- Discretización de variables continuas