

Transformación de Datos

Sesión 01

Ing. Gómez Marín, Jaime¹

Módulo 3 : Análisis de Datos con Python
Departamento de TD



- Introducción
- Motivación
- KDD : Proceso de extracción del Conocimiento
- Metodologia CRISP-DM
- Preparar los datos
- Datos perdidos : Impacto
- Conclusiones
- Bibliografía

¿ Porque extraer datos ?

- Gran cantidad de datos coleccionada y almacenada
- Las computadoras se vuelven más baratas y poderosas
- La presión de la competencia es fuerte
- Data recolectada y almacenada a enorme velocidad (GB/hora).



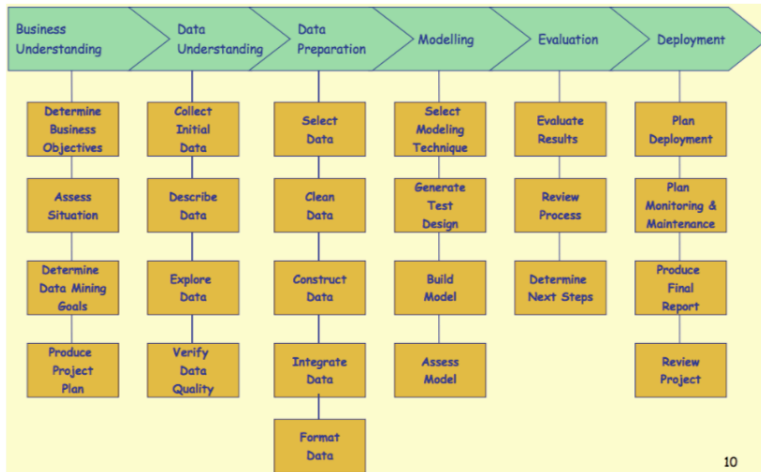
Las Herramientas y nuevas tecnologías automatizadas de almacenamiento de datos conllevan a que grandes cantidades de datos sean almacenados en bases de datos, data warehouses y otros repositorios de información.



CRISP-DM es una metodología completa de Minería de Datos y modelamiento de procesos que proporciona a cualquiera, desde principiantes a expertos, un programa completo para la realización de un proyecto de minería de datos. La metodología enumera los pasos para reproducir el éxito



CRISP-DM : Fases



El propósito de la preparación es transformar los conjuntos de datos de tal forma que la información que contienen esté mejor expuesta para la herramienta de minería de datos que se utilizará.



- Identificar y manejar valores perdidos
 - Identificar valores perdidos
 - Evaluar datos perdidos
 - Corregir el formato de los datos
- Estandarizar datos
- Normalización de datos
- Binning

Impacto de los valores faltantes:

- 1 % datos faltantes – trivial.
- 1-5 % – manejable
- 5-15 % – requiere métodos sofisticados
- Más del 15 % – interpretación perjudicial

En esta sesión se han usado la forma de como poder obtener y limpiar los datos usando tecnicas de Minería de Datos.



Naomi Ceder. The Quick Python Book - Manning Publications, 2018.