

***¡LES DAMOS LA
BIENVENIDA!***

¿Comenzamos?



Semana 02. ANÁLISIS Y GESTIÓN DE DATOS

***MÓDULO ANALISTA DE
BIG DATA***



OBJETIVOS DE LA CLASE

- Procesamiento de datos masivo.
- Metodologías de procesamiento.



INTRODUCCIÓN AL PROCESAMIENTO DE DATOS MASIVOS

¿A QUÉ LLAMAMOS PROCESAMIENTO DE DATOS?

El almacenamiento y transformación de
elementos de **datos** para producir
información significativa.

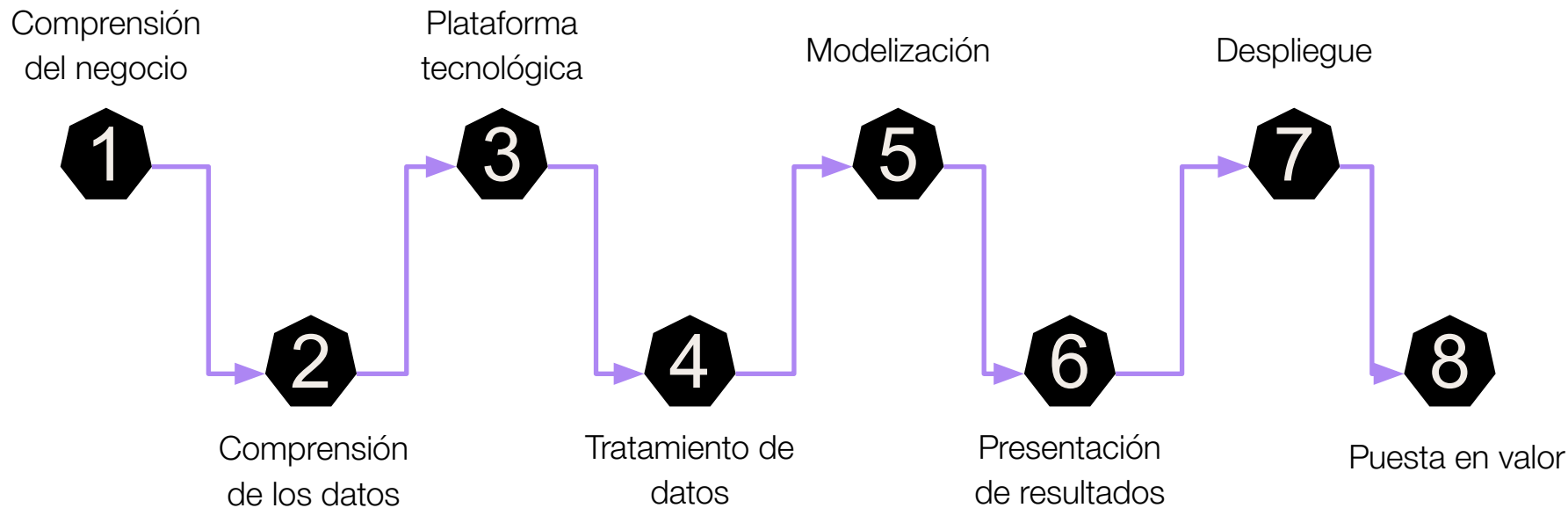
METODOLOGÍAS

La metodología de procesamiento de grandes volúmenes de datos permite transformar los datos en conocimiento. Para ello, vamos a ver tres grandes cosas:

1. Primero, la propia metodología de procesamiento de datos;
2. Segundo, cuáles son las componentes de esa metodología en alto nivel;
3. Y por último, los factores de éxito para que esta metodología sea una realidad en las compañías.

METODOLOGÍA DE PROCESAMIENTO

Diagrama de las 8 fases



COMPONENTES

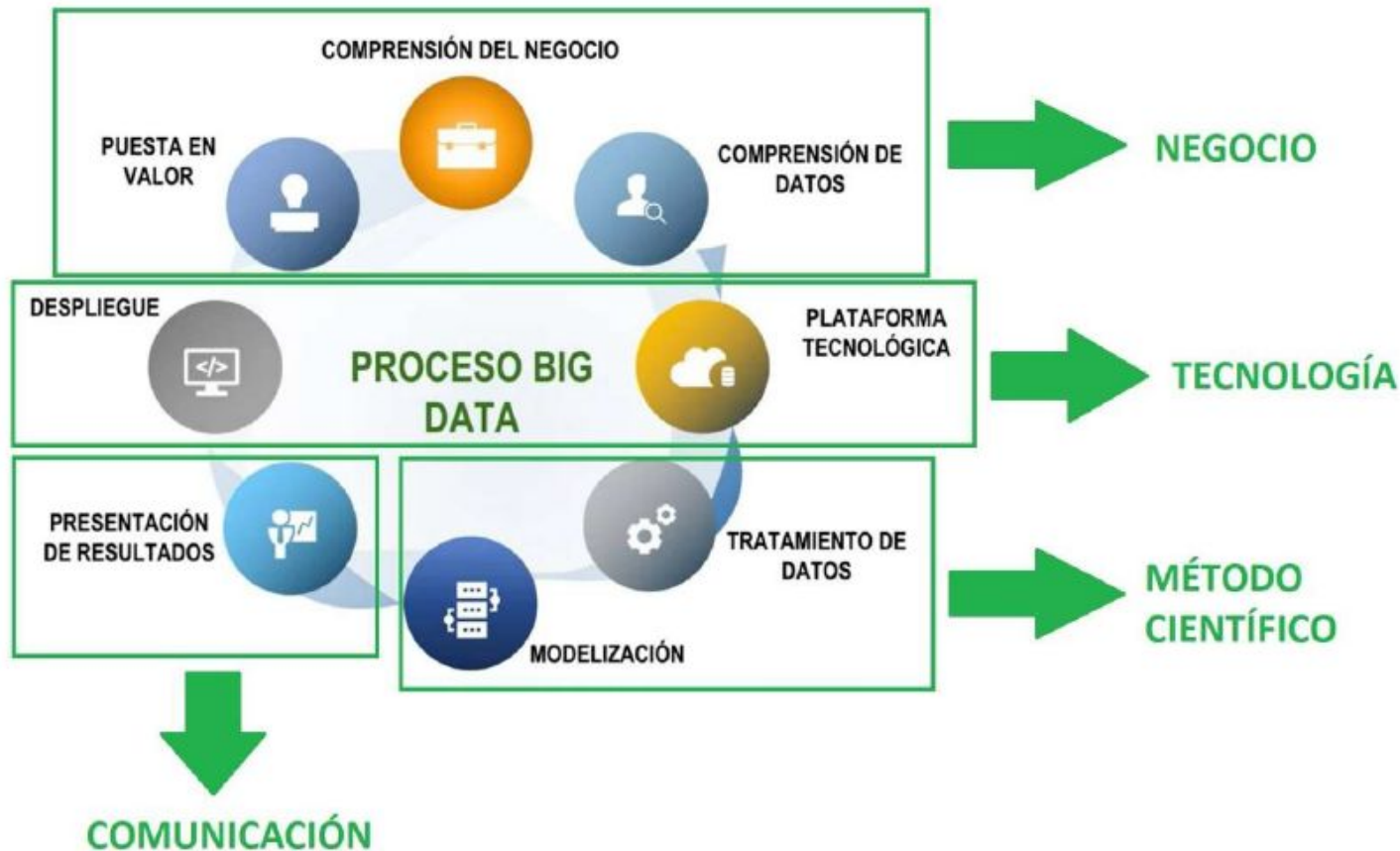
Componente del negocio: en un modelo lo importante es que sea accionable, un modelo tiene que resolver un problema real

Componente tecnológica: Sin esta tecnología, no vamos a ser capaces de procesar grandes volúmenes de información

Componente científica: construyen aplicaciones utilizando el método científico, utilizando técnicas analíticas.

Componente comunicación: transmitir resultados.

Metodología de Procesamiento





***¿QUÉ NECESITAMOS PARA
IMPLEMENTAR ESTA
METODOLOGÍA?***

COMPONENTES

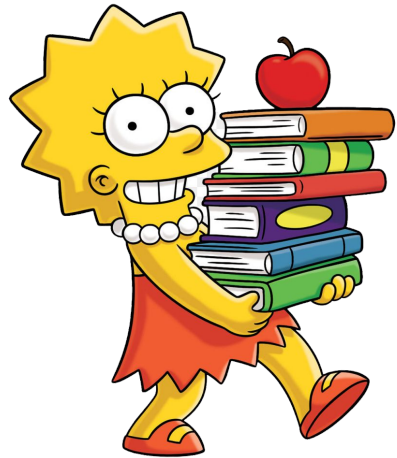
DATOS: Sin datos no podemos hacer nada

TALENTO: conocimientos adecuados y las capacidades adecuadas
para poder tratar toda esta información.

COMPONENTES

HERRAMIENTAS ANALÍTICAS Y TECNOLOGÍA: Si tenemos datos, pero no somos capaces de tratarlos, no somos capaces de tener una plataforma tecnológica suficientemente potente para poder desarrollar esos modelos analíticos, no vamos a poder terminar el trabajo.

CULTURA ORGANIZACIONAL: hace falta una cultura organizacional en la que se premie toda esta visión de negocio y, sobre todo, se entienda que la ciencia viene a aportar valor.



***VEAMOS EN PROFUNDIDAD CADA UNA DE LAS METODOLOGÍAS
QUE NOMBRAMOS ANTERIORMENTE... ¿LES PARECE?***

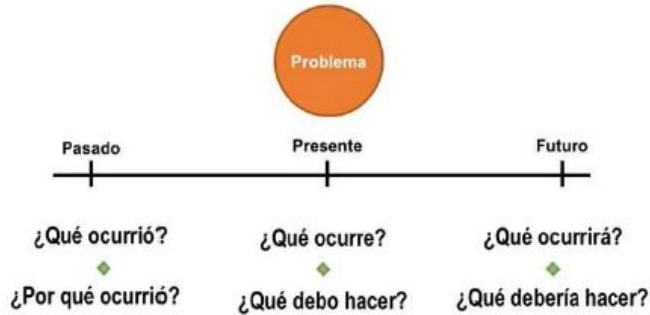
Comprensión del negocio

OBJETIVO: Identificar, analizar y comprender el problema y traducirlo a un problema analítico.

ETAPAS:

- Identificación del problema.
- Fijación de los objetivos.
- Identificación de los implicados.
- Fijación de la tipología de análisis.

Objetivo

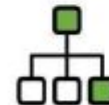


Es muy importante que tengamos claro el objetivo porque cada objetivo tiene una técnica analítica distinta y el tratamiento de información es diferente.

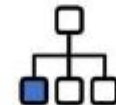
Tendremos que entender a quién impacta este problema, quién nos puede aportar conocimiento y quién lo va a utilizar finalmente.

Áreas Implicadas

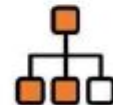
Tecnológicas



Analíticas



Negocio



Es fundamental comprender el problema de negocio de forma analítica.

Problema Analítico Subyacente

Análisis Descriptivo

- Mostrar mediante estadísticos la realidad capturada

Análisis Inferencial

- Generalizar conclusiones muestrales a toda la población, estudiar las relaciones entre variables y contrastar hipótesis

Análisis Predictivo

- Determinar datos futuros a través de datos históricos

Análisis Prescriptivo

- Recomendar la acción adecuada y sus consecuencias

Retos

Los principales **retos** de esta fase son:

- ✓ No identificar algún interviniente
- ✓ Establecer un objetivo claro
- ✓ Fijar las expectativas
- ✓ Menospreciar el conocimiento específico del problema a afrontar

Sin **Comprensión de Negocio** se incrementa el riesgo de construir modelos que no aporten valor

Comprensión de los datos

OBJETIVO: Identificar las fuentes de información y analizar su conveniencia para su posterior captura y almacenamiento.

ETAPAS:

- Inventario de información.
- Identificación de las fuentes.
- Disponibilidad de las fuentes.
- Relación de la información.
- Representación funcional de datos.

¿Dónde encontramos esta información?

Fuentes de Información: Identificación

Fuentes
Internas



Redes
Sociales

Fuentes
Externas

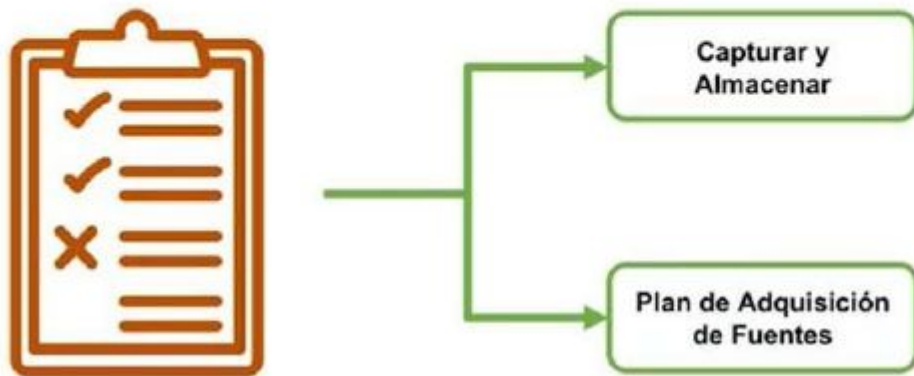


Open
Data

¿Dónde se encuentra la
información?

¿Es accesible esta información?

Fuentes de Información: Disponibilidad



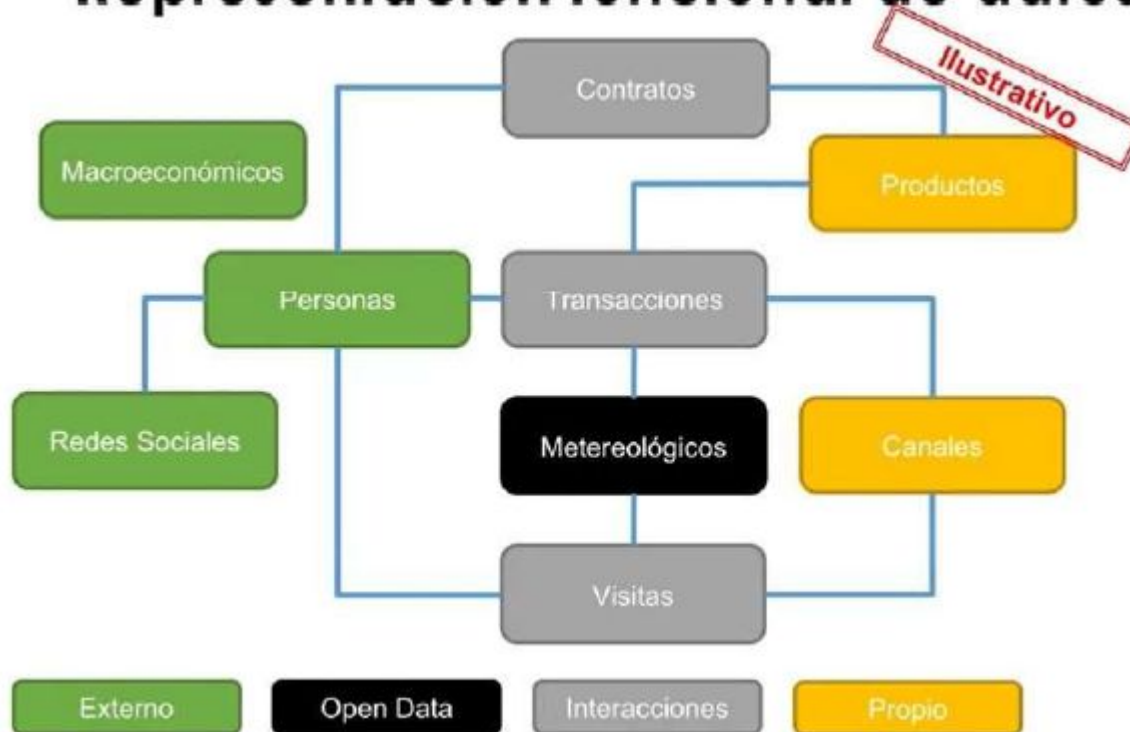
¿Podemos relacionar los conceptos?

Relación de la Información



¿Cuál es la mejor forma de representarlos?

Representación funcional de datos



Retos

Los principales retos de esta fase son:

- ✓ Identificar las fuentes de Información asociadas al problema de negocio
- ✓ Comprender la información contenida en los datos
- ✓ Relacionar los conceptos
- ✓ No focalizarse en los datos disponibles

La Comprensión de Datos aflora el contexto de los datos y sus relaciones

Plataforma tecnológica

OBJETIVO: Disponer de una plataforma tecnológica para la construcción del modelo analítico.

ETAPAS:

- Diseño de la arquitectura tecnológica.
- Selección de componentes Big Data.
- Estrategia de implantación.

A la hora de definir el diseño de la arquitectura tecnológica debemos:

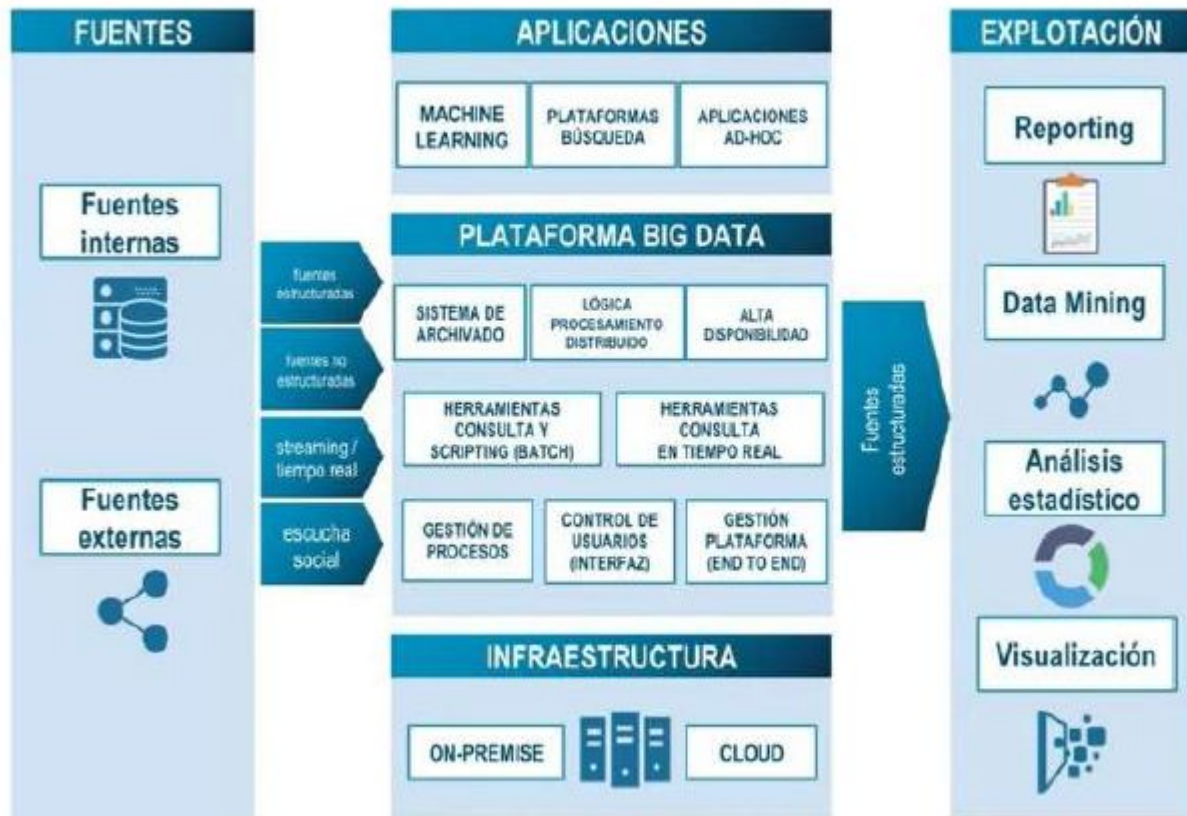
Diseño Arquitectura Tecnológica

Establecer la **composición** de los **elementos estructurales** necesarios para soportar las acciones asociadas a la construcción y explotación del modelo analítico:

- **Captura de datos**
- **Almacenamiento de datos**
- **Procesado de datos**
- **Explotación del modelo**



Diseño Arquitectura Tecnológica



Componentes Big Data

Comparación tecnológica

2012 Big Data Landscape



Estrategía de Implantación



On-Premise

IaaS
Infraestructura

PaaS
Plataforma

SaaS
Software

Datos	Datos	Datos	Datos
Aplicaciones	Aplicaciones	Aplicaciones	Aplicaciones
Ejecución	Ejecución	Ejecución	Ejecución
Bases de Datos	Bases de Datos	Bases de Datos	Bases de Datos
Servidores	Servidores	Servidores	Servidores
Virtualización	Virtualización	Virtualización	Virtualización
Almacenamiento	Almacenamiento	Almacenamiento	Almacenamiento
Comunicaciones	Comunicaciones	Comunicaciones	Comunicaciones

Responsable

Usuario

Proveedor

Retos

Los principales **retos** de esta fase son:

- ✓ Considerar todas las implicaciones de la arquitectura diseñada
- ✓ Estar al día de la evolución de las componentes Big Data y su interrelación
- ✓ Dimensionar de forma adecuada los recursos tecnológicos necesarios
- ✓ Establecer una estrategia adecuada

Desplegar una **Plataforma Tecnológica** errónea puede suponer la imposibilidad de construcción y despliegue del modelo analítico

Tratamiento de datos: Preparación

OBJETIVO: Capturar, almacenar y preparar la información.

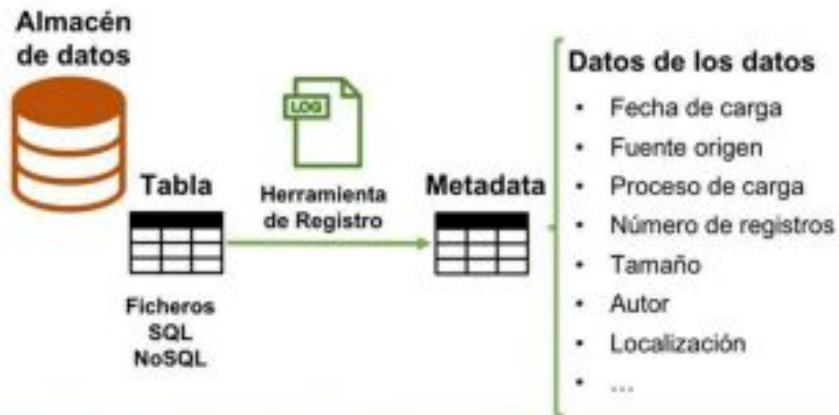
ETAPAS:

- Adquisición y registro.
- Metadatado.
- Exploración y análisis.
- Calidad de dato y limpieza.

Motivación



Metadatado de Tablas

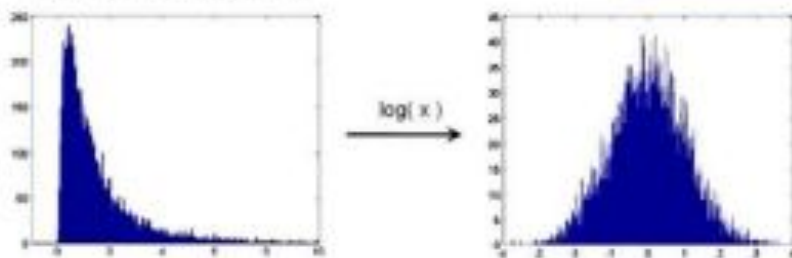


El Metadatado es clave para un Gobierno del Dato adecuado

Formateo y Construcción de Variables

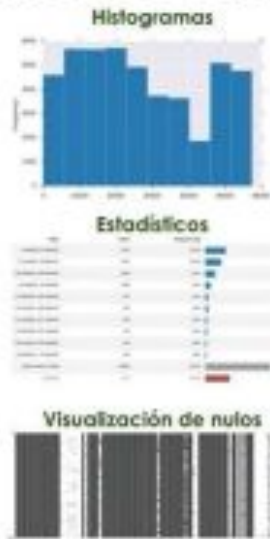


Transformación de variables:



Exploración y Análisis de Variables

Tabla

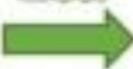


Calidad del dato y Limpieza

Tabla



Herramienta
de Calidad



Limpieza

- Tratar campos vacíos
- Tratar valores incoherentes
- Analizar valores atípicos (outliers)
- Corregir errores
- Eliminar repetidos
- Normalizar datos
- ...

Retos

Los principales retos de esta fase son:

- ✓ Evaluación de la calidad de los datos
- ✓ Tratamiento de información no estructurada
- ✓ Fijación de criterios de tratamiento
- ✓ Diseño de la política de metadatos

La Preparación de Datos asegura disponer de datos de calidad que permitan extraer el conocimiento

Tratamiento de datos: Fusión

OBJETIVO: Construir un tablón único de datos con toda la información disponible

ETAPAS:

- Representación de datos.
- Análisis de integridad.
- Integración de tablas.
- Construcción de variables derivadas.

Representación de los datos



Esquema de Bases de Datos

Análisis de Integridad



PK-Clave Primaria

Integridad de entidad

FK-Clave Foránea

Integridad referencial

Integración de tablas



Construcción de Variables Derivadas



Retos

Los principales retos de esta fase son:

- ✓ Diseño del modelo de datos
- ✓ Evolución del modelo de datos
- ✓ Gestión de las incidencias en la integración de tablas

La **Fusión de Datos** permite relacionar todos los conceptos asociados al problema

Modelización

OBJETIVO: Construir un modelo analítico.

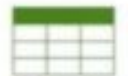
ETAPAS:

- Tipología de técnicas de modelado.
- Diseño de técnicas de modelado.
- Diseño de técnicas de evaluación.
- Entrenamiento del modelo.
- Evaluación del modelo.

Tipologías de Técnicas de Modelado



Diseño de Técnicas de Modelado



Datos



Modelo



Función de
coste



Algoritmo
estimador

x_1	x_m	y

Ilustrativo

$$y = a + b_1x_1 + \dots + b_mx_m$$

$$g(a, b_1, \dots, b_m) = \sum_{i=1}^n (y_i - a - b_1x_{i1} - \dots - b_mx_{im})^2$$

Minimos Cuadrados Ordinarios,
Descenso del Gradiente

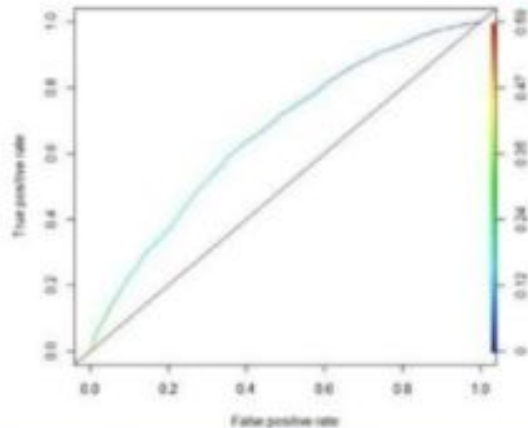
Cada Técnica tiene sus propias
componentes

Diseño de la Evaluación



AUC_ROC
AIC
BIC
RMSE
KS

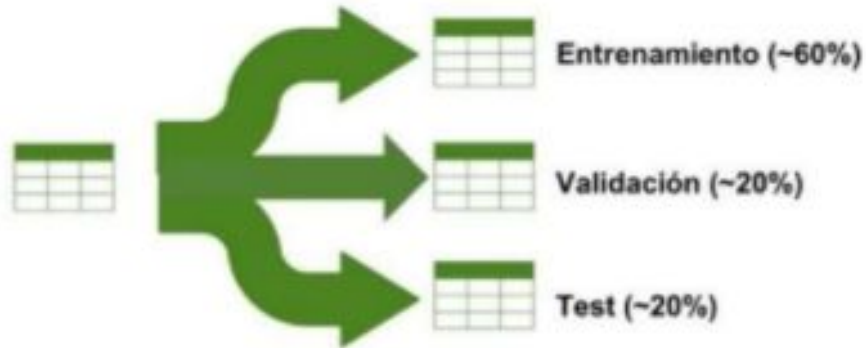
...



Ilustrativo

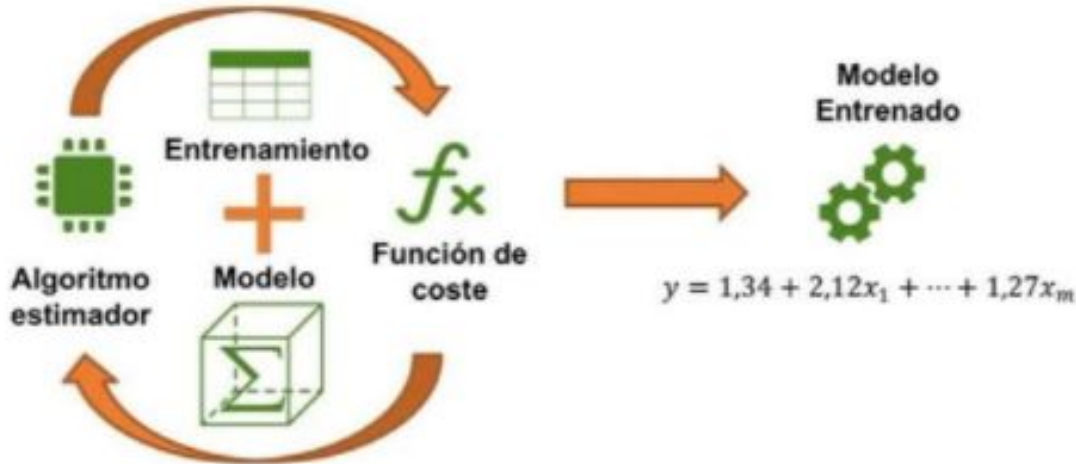
El criterio de evaluación debe ser
claro y único

Entrenamiento del Modelo: División



Asegura la capacidad analítica y evita el overfitting

Entrenamiento del Modelo: Estimación



Se entrenan los modelos con el
conjunto de entrenamiento

Entrenamiento del Modelo: Selección



Se selecciona el modelo con el
conjunto de validación

Evaluación del Modelo



La capacidad analítica se calcula en el **conjunto de test**

Retos

Los principales retos de esta fase son:

- ✓ Decidir entre cajas blancas y cajas negras
- ✓ Seleccionar una métrica de evaluación adecuada al problema
- ✓ Evitar los problemas de overfitting
- ✓ Equilibrar el modelo entre bias y variance
- ✓ Disponer de suficientes datos para la convergencia los algoritmos estimadores
- ✓ Disponer de suficiente potencia de cálculo para la convergencia de los algoritmos estimadores

En la **Modelización** se construyen modelos, simplificaciones de la realidad, que ayudan a comprender lo complejo de forma sencilla

Presentación de resultados

OBJETIVO: Trasladar el conocimiento al resto de los intervinientes implicados

ETAPAS:

- Informes y reportes.
- Visualizaciones.
- Infografías.
- Cuadros de mando.

Informes y Reportes

Simplificación de información mediante estadísticos y gráficos

Auditoría de la tabla "Aterrizaje-Meteoritos"

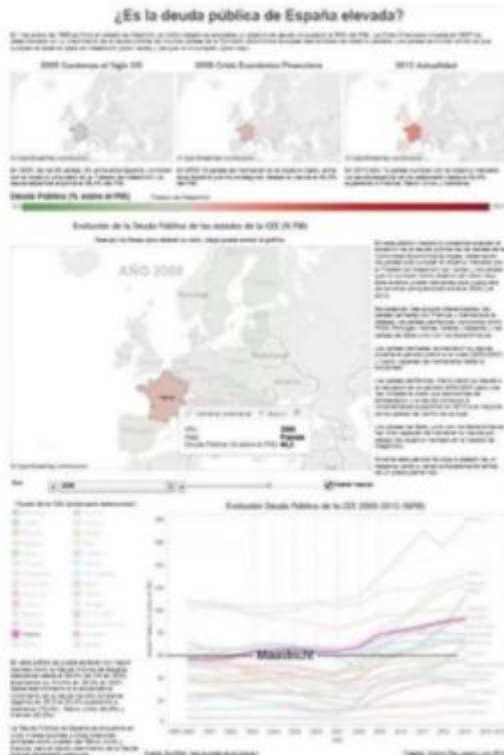
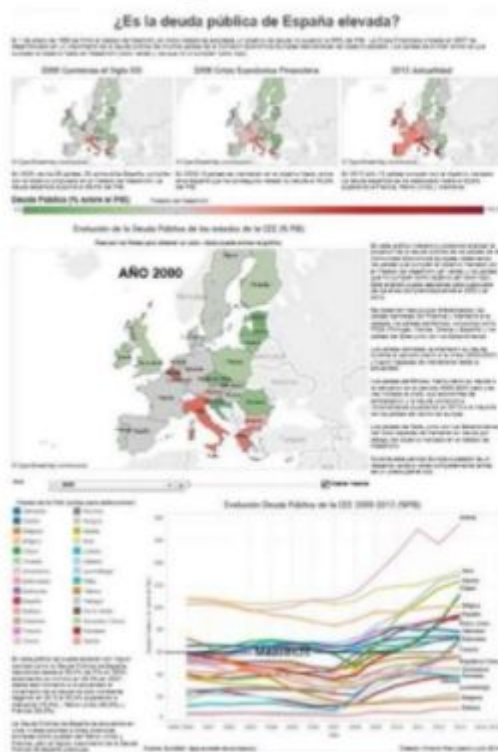
Descripción de los datos de la tabla:

Numero de variables (columnas) en la tabla:	10
Numero de registros en la tabla:	45716
Porcentaje de valores que no están imputados:	4.9%
Variables (columnas) de tipo numérico:	4
Variables (columnas) de tipo categórico:	5
Variables en formato fecha:	0
Variables de tipo texto con valor único:	1




Alertas

- La variable `Geolocation` tiene 7315 registros no imputados lo que supone un 16.0% del total **No imputados**
- La variable `Geolocation` tiene una alta cardinalidad con 17101 valores distintos **Alerta**
- La variable `sex` está muy sesgada (1 = 76.90%)

Infografías Interactivas

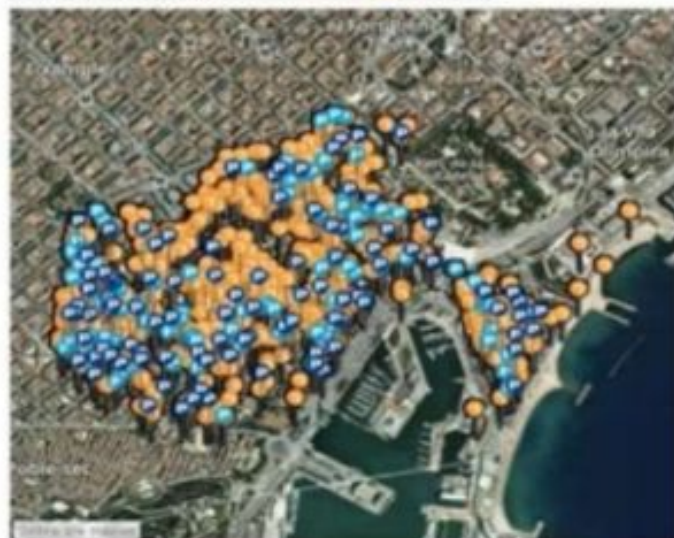


Clients / No Clients	Municipality	District	Neighbourhood	ID_ZONA	ID_AREA	ID_DISTRITO	Ver Tabla Completa
(Todos)	BARCELONA	Ciutat Vella	(Todos)	(Todos)	(Todos)	(Todos)	

Clients Without Sales Potential  99
 Clients With Sales Potential  101
 Potential New Clients  799

ID: 570996 - SUPERMERCADO

PAU CLARIS Nº 74, BARCELONA Tel: 670642952
 Distrito: Eixample Barrio: la Dreta de l'Eixample
 Zona: 1 Area: 1 Distrito: 1



Sales

1,4 M €

Market Share

33%

Total Store Sales

4,2 M €



Neighborhood Classification

IN HOME  
 KIDS  
 ON THE GO  
 SHARING  
 TOURIST  

Current Category



Potential Category



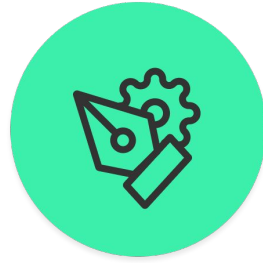
Trade Marketing
 Execution
 Tourist

Retos

Los principales retos de esta fase son:

- ✓ Adecuar la presentación al nivel de los intervinientes
- ✓ Transmitir los mensajes adecuados
- ✓ Mostrar lo importante y no lo interesante

La **Presentación de Resultados** es el momento clave en el que se trasmite el conocimiento al resto de intervinientes



TRABAJO EN LA PLATAFORMA

¿PREGUNTAS?



¡MUCHAS GRACIAS!