

Ciencia de Datos Introducción

Master in Business Administration

Andrés Martínez

31 Mayo, 2024

- Logística del Módulo.
- Qué es Ciencia de Datos?
- Surgimiento de la Ciencia de Datos.
- Aplicación a nivel gerencial.
- Metodología de la Ciencia de Datos.
- Deductivo e Inductivo.
- Inferencia.
- Trade off en Ciencia de Datos.

- 4 horas por módulo.
- Usamos Orange, Power Bi, Excel y Python.
- Proyecto 30%.
- Casos de Estudio 30%.
- Prueba de selección múltiple 10%.

Primer Día

- Introducción
- Metodología

Tercer Día

- Visualización
- Data Story Telling

Quinto Día

- Introducción a Python
- Bases de Datos

Segundo Día

- Desarrollo del problema
- Configuración KPI

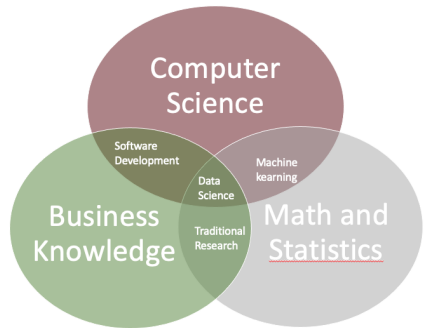
Cuarto Día

- Introducción a Orange
- Metodología en Orange

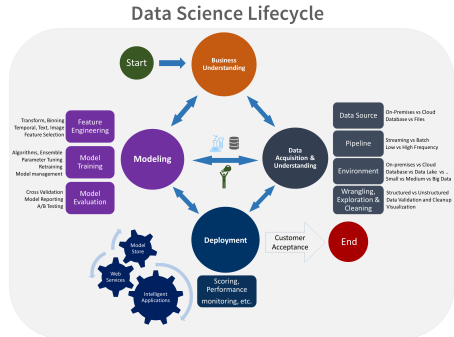
Sexto Día

- Presentación de proyectos
- Conclusión

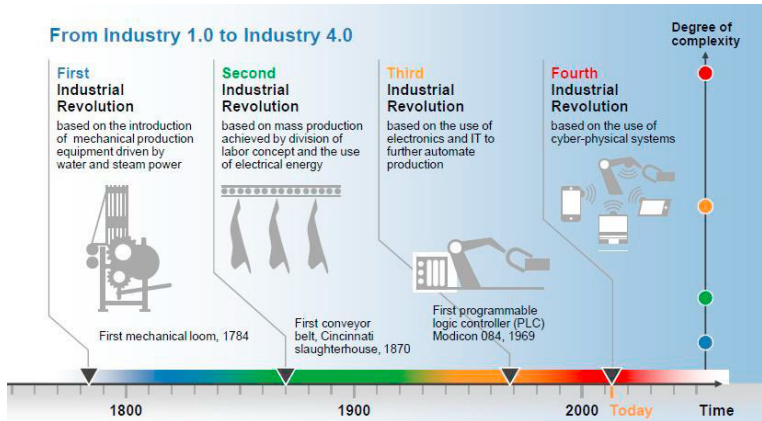
- "Data science is the scientific practice of extracting knowledge from data, usually referring to large and complex bodies of data" Pietsch (2022, p. 2).
- "Data science could be defined as an interdisciplinary knowledge to understand the real world, using massive amounts of data to create technological artifacts which provide information to make decisions in different fields" Andrés Martínez.



- La Ciencia de Datos es un método que logra conectar diferentes niveles de la organización a través de la interpretación de las métricas de los modelos de predicción y clasificación y la capacidad que tiene el responsable de las decisiones de conectar los resultados con los KPI.



History Matters



Estadística Tradicional vs. Ciencia de Datos

Estadística Tradicional (Método Deductivo)

- Enfoque deductivo.
- Define hipótesis a priori.
- Selecciona técnicas estadísticas.
- Recopila datos específicos.
- Prueba hipótesis.
- Conclusiones basadas en hipótesis.
- Emplea muestras representativas.

Ciencia de Datos (Método Inductivo)

- Enfoque inductivo.
- Explora datos sin prejuicios.
- Recopila grandes volúmenes de datos.
- Descubre patrones y relaciones.
- Conclusiones emergen de los datos.
- No depende de muestras representativas.

Domain Knowledge KPI



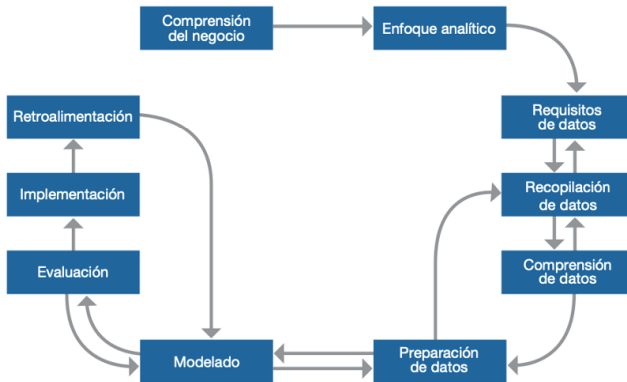


Figure 1: IBM

Metodologías en Ciencia de Datos

CRISP-DM

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Implementación

SEMMA (Sample, Explore, Modify, Model, Assess)

- Muestra (Sample)
- Exploración (Explore)
- Modificación (Modify)
- Modelado (Model)
- Evaluación (Assess)

KDD (Knowledge Discovery in Databases)

- Selección
- Preprocesamiento
- Transformación
- Minería de datos
- Interpretación/Evaluación

Metodología Ágil

- Iterativa
- Enfoque en colaboración y flexibilidad
- Desarrollo incremental
- Retroalimentación continua
- Adaptación rápida a cambios

1. Comprensión del Negocio

- Definir objetivos y requisitos del negocio.
- Convertir objetivos de negocio en objetivos de minería de datos.

2. Comprensión de los Datos

- Recolectar datos iniciales.
- Describir los datos.
- Explorar los datos.
- Verificar la calidad de los datos.

3. Preparación de los Datos

- Seleccionar datos relevantes.
- Limpiar los datos.
- Construir datos.
- Integrar datos.
- Formatear datos.

4. Modelado

- Seleccionar técnicas de modelado.
- Diseñar pruebas.
- Construir modelos.
- Evaluar modelos.

En los últimos años, el concepto de Economía Circular (CE) ha recibido una atención significativa como una forma de promover el desarrollo sostenible. La CE busca desvincular la creación de valor del consumo de recursos finitos mediante estrategias que mantengan productos, componentes y materiales en uso por más tiempo. La adopción de la CE por la industria aún es modesta, y aquí es donde las Tecnologías Digitales (DTs) aportar en el desarrollo sostenible.

Tecnologías Digitales (DTs)

- Internet de las Cosas (IoT)
- Big Data
- Inteligencia Artificial (AI)

Estas DTs pueden ser la base para la adopción y aceleración de la transición a la CE, formando los bloques operativos de una CE más eficiente y efectiva, conocida como Smart CE.

Para apoyar la integración efectiva de la ciencia de datos dentro de las organizaciones en el contexto de la CE, se propone un modelo de proceso CRISP-DM mejorado con las siguientes modificaciones:

1. Fase de Validación de Datos

- Requiere la re-involucración de la entidad empresarial para validar que los datos preparados representan adecuadamente el problema original.

2. Perfiles Analíticos

- Estructuras que estandarizan la colección, aplicación y reutilización de conocimientos analíticos y modelos para entidades clave del negocio.

Estas modificaciones abordan la falta de retroalimentación al nivel empresarial y la falta de control del valor añadido.

Mantenimiento Predictivo (PdM) en el Contexto de la Economía Circular

- **Definición:** El mantenimiento predictivo se define como un mantenimiento basado en condiciones que se lleva a cabo siguiendo una previsión de análisis o características conocidas de las características de degradación de un activo.
- **Contraste con Mantenimiento Tradicional:** A diferencia del mantenimiento tradicional, que solo se basa en la información de la condición actual.
- **Integración de Tecnologías Digitales:** El PdM integra múltiples tecnologías digitales (e.g., Internet de las Cosas e Inteligencia Artificial), permitiendo el acceso en tiempo real a información detallada sobre la ubicación, condición y disponibilidad de los activos.
- **Ventajas:**
 - Mejora la toma de decisiones al predecir la salud, desgaste, uso y consumo de energía de los productos.
 - Aumenta la transparencia de la condición actual de los activos durante su ciclo de vida.
 - Facilita la activación de operaciones adecuadas para extender el ciclo de vida de los activos para el fabricante de equipos originales (OEM)

Mantenimiento Predictivo (PdM) para un Fabricante de Equipos Originales (OEM)

Objetivo

- Extender el ciclo de vida de los activos.
- Aumentar la utilización de los equipos.
- Mejorar la transparencia en la condición y el historial de uso de los activos.

Resultados

- Evaluación de la preparación de datos y validación.
- Identificación de anomalías y desarrollo de métodos para evaluar el grado de severidad.

El estudio de caso mostró que la validación de datos y el uso de perfiles analíticos mejoran la vista de gestión y la comunicación del valor empresarial.

1. Selección

- Seleccionar datos relevantes para la tarea de análisis.

2. Preprocesamiento

- Limpiar y filtrar datos para eliminar ruido e inconsistencias.

3. Transformación

- Transformar los datos en un formato adecuado para la minería de datos.

4. Minería de Datos

- Aplicar técnicas de minería de datos para extraer patrones y modelos.

5. Interpretación/Evaluación

- Evaluar la utilidad de los patrones descubiertos.
- Interpretar y visualizar los resultados.

1. Muestra (Sample)

- Extraer una muestra representativa de los datos.

2. Exploración (Explore)

- Explorar los datos para encontrar patrones iniciales.
- Visualizar los datos.

3. Modificación (Modify)

- Crear, seleccionar y transformar variables para enfocarse en la modelización.

4. Modelado (Model)

- Aplicar técnicas de modelización para predecir resultados.

5. Evaluación (Assess)

- Evaluar la calidad y efectividad del modelo.

1. Planificación

- Definir el alcance del proyecto y los objetivos principales.

2. Desarrollo Iterativo

- Desarrollar el proyecto en ciclos cortos con incrementos frecuentes.

3. Retroalimentación Continua

- Recibir y actuar sobre la retroalimentación continua de los stakeholders.

4. Adaptación Rápida

- Adaptar rápidamente el proyecto según sea necesario basado en la retroalimentación y cambios en el entorno.

Ventajas y Desventajas de las Metodologías

CRISP-DM

- Ventajas:
 - Estructura clara y definida.
 - Flexible y adaptable.
- Desventajas:
 - Puede ser complejo de implementar en proyectos pequeños.
 - Requiere una buena comprensión del negocio.

KDD

- Ventajas:
 - Enfoque claro en la minería de datos.
 - Bien documentado y estandarizado.
- Desventajas:
 - Menos flexible que CRISP-DM.
 - Puede ser difícil de adaptar a proyectos no estándar.

Ventajas y Desventajas de las Metodologías

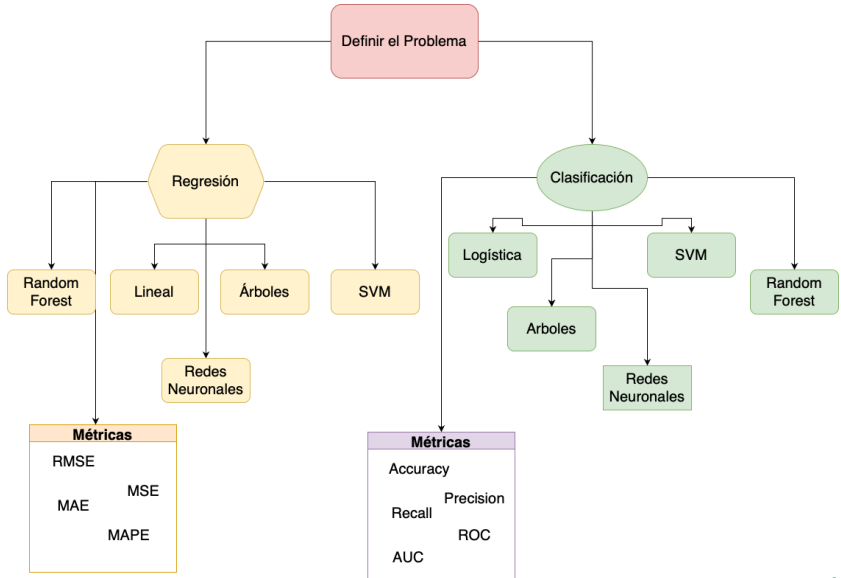
SEMMA

- Ventajas:
 - Enfoque en el modelado y la evaluación.
 - Buen soporte para la exploración de datos.
- Desventajas:
 - Falta de enfoque en la comprensión del negocio.
 - Más adecuado para proyectos de SAS.

Metodología Ágil

- Ventajas:
 - Alta flexibilidad y adaptabilidad.
 - Fuerte énfasis en la colaboración y la retroalimentación.
- Desventajas:
 - Puede ser difícil de gestionar sin experiencia previa.
 - Requiere cambios culturales en la organización.

Modelos Supervisados



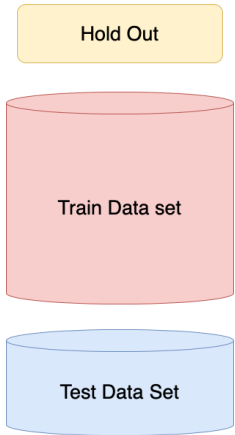


Figure 2: Hold out

The diagram illustrates the 'Cross Validation' method for data partitioning. At the top, a red rounded rectangle is labeled 'Cross Validation'. Below it is a 6x6 grid of cells. Each cell contains either the word 'Train' or 'Test'. In each row and each column, exactly one cell contains the word 'Test' and is highlighted in light blue, while the other five cells contain the word 'Train' and are highlighted in light green. This represents 6 different iterations of training and testing on different subsets of the data.

Train	Train	Train	Train	Train	Test
Train	Train	Train	Train	Test	Train
Train	Train	Train	Test	Train	Train
Train	Train	Test	Train	Train	Train
Train	Test	Train	Train	Train	Train
Test	Train	Train	Train	Train	Train

Figure 3: Cross Validation

Tradeoff

