Ciencia de Datos Introdución

Certificación en Ciencia de Datos

Andrés Martínez

31 Mayo, 2024

Contenido

- 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.

Logística

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

Introducción

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

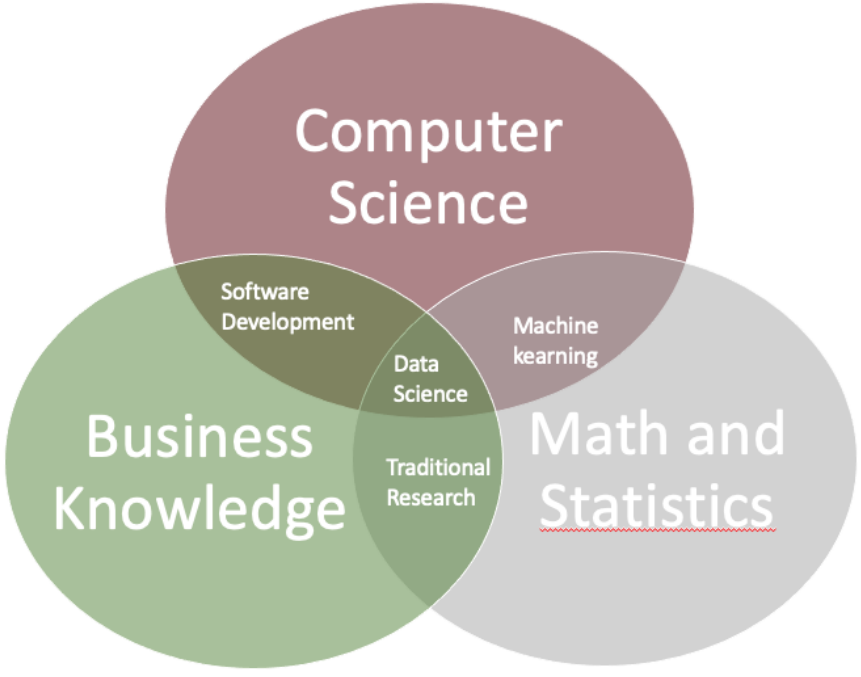
- Introducción a Orange
- Metodología en Orange

Sexto Día

- Presentación de proyectos
- Conclusión

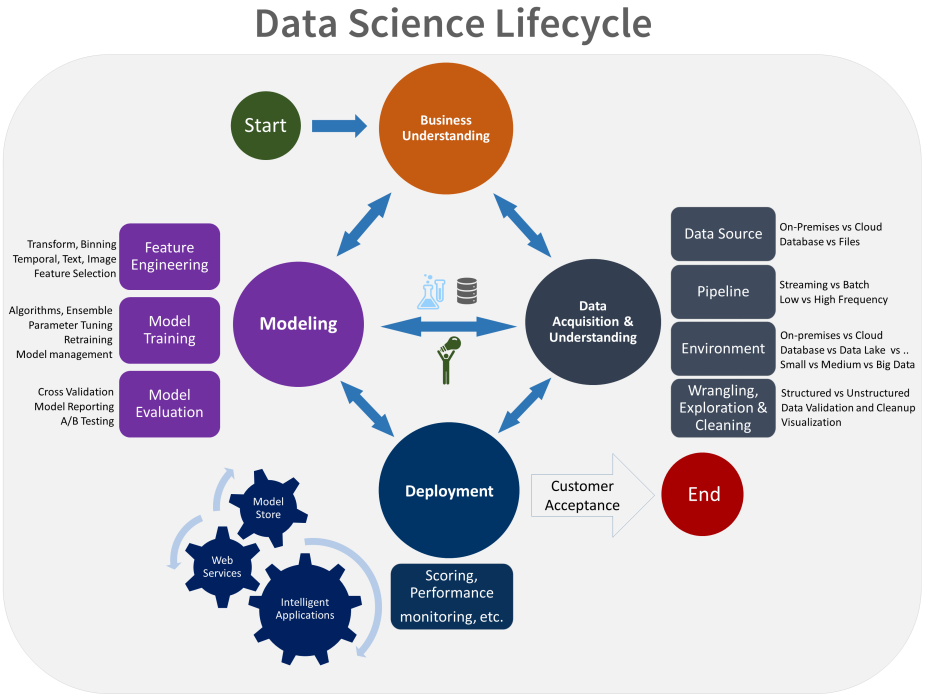
Ciencia de Datos

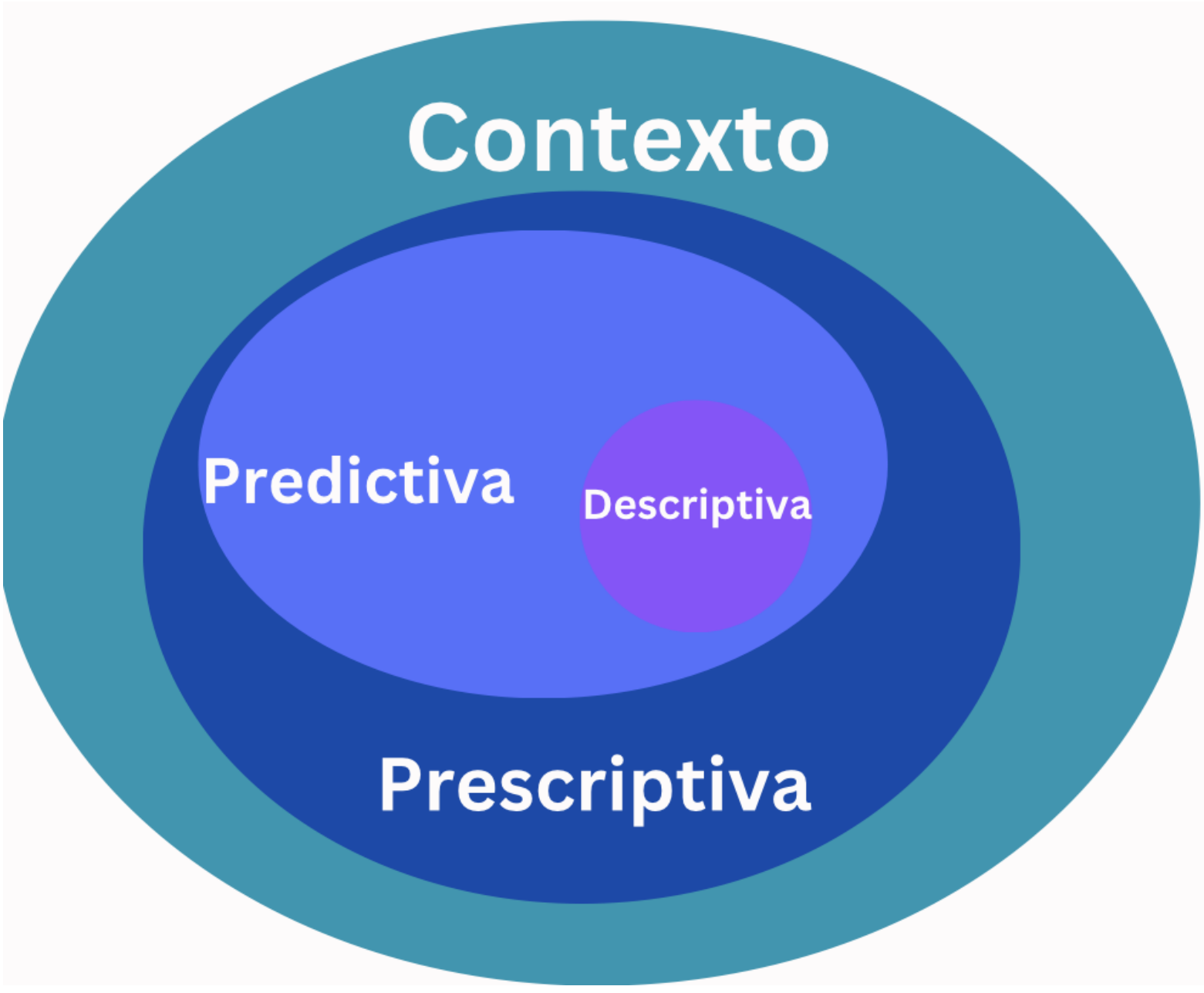
- "Data science is the scientific practice of extracting knowledge from data, usually referring to large and complex bodies of data" Pietsch (2022, p. 2).
- "Data science could be defined as an interdisciplinary knowledge to understand the real world, using massive amounts of data to create technological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



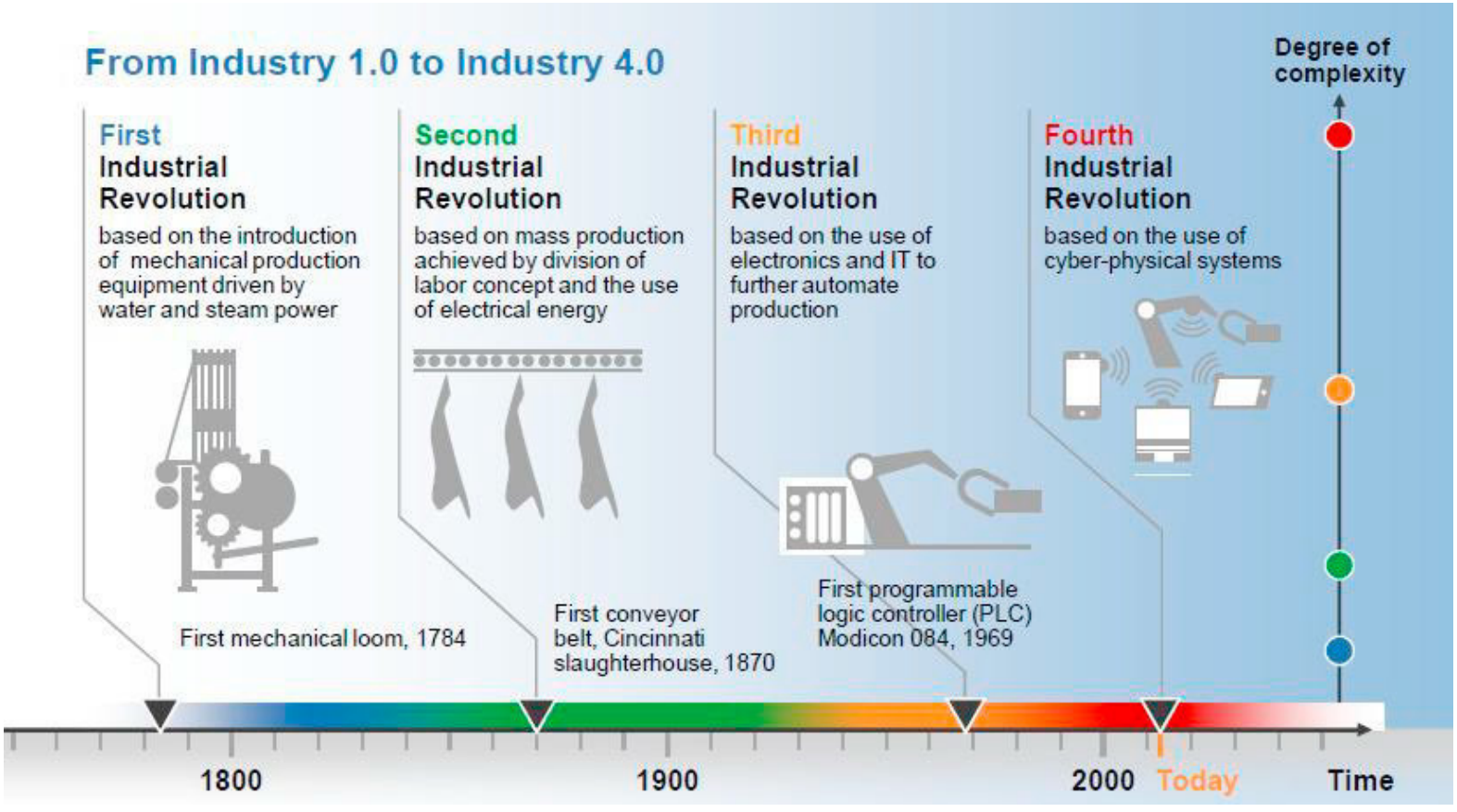
Ciencia de Datos en Gestión

- 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.




Contexto
Predictiva
Descriptiva
Prescriptiva

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

Métricas

Cuáles son las métricas más sensibles del proyecto

Preguntas

Determinar las preguntas iniciales y como afectan el resultado esperado

Datos

Fuentes de recolección de datos y sus conexiones con el problema

Necesidades

Ojo con el síndrome de Adam, hay proyectos que ya se hicieron algunos fueron exitosos y otros fallaron

Factores

Qué se puede controlar y que no se puede controlar para lograr el objetivo

Metodología

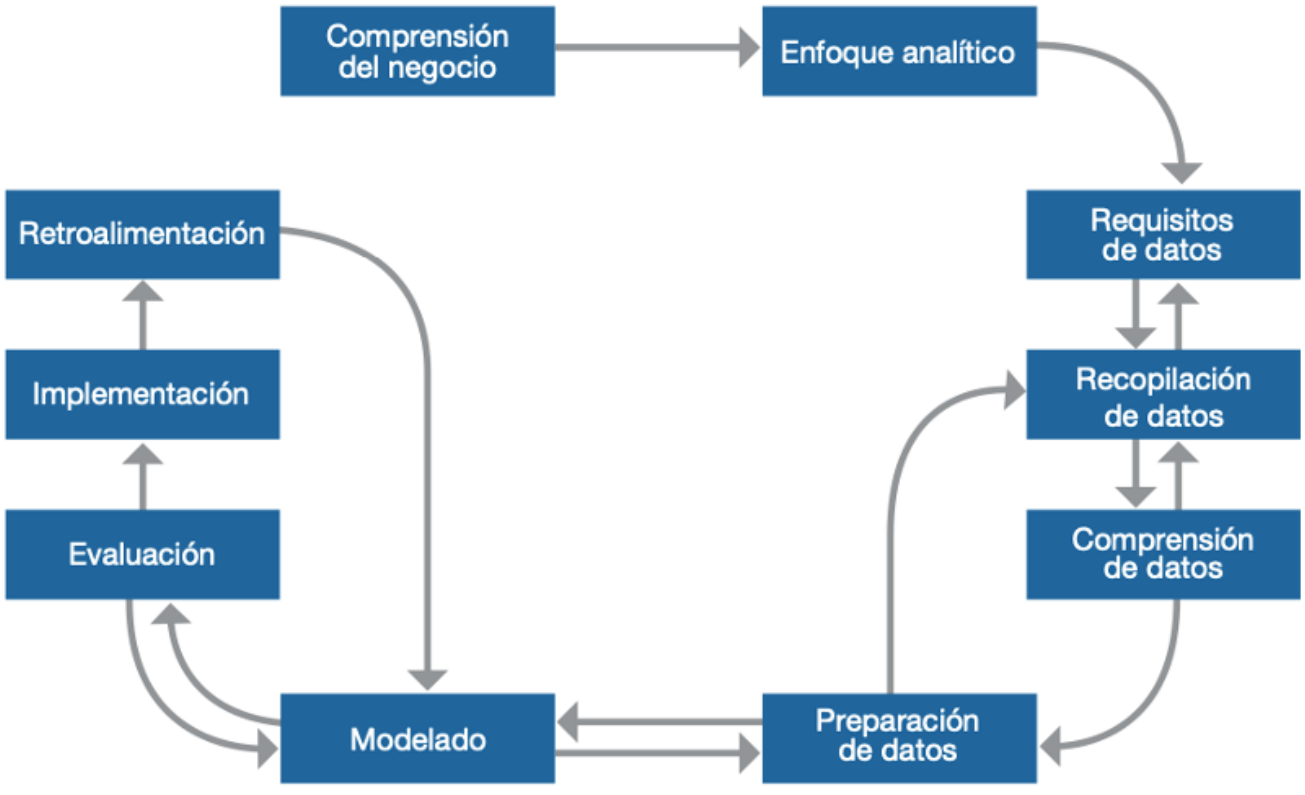


Figure 1: IBM

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

CRISP-DM: Fases Iniciales

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.

CRISP-DM: Fases Posteriores

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.

CRISP-DM y la Economía Circular: Introducción

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) aportan 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.

CRISP-DM y la Economía Circular: Modificaciones

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)

CRISP-DM y la Economía Circular: Estudio de Caso

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 anormalidades 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.

KDD: Fases Iniciales

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.

KDD: Fases Posteriores

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.

SEMMA: Fases Iniciales

1. **Muestra (Sample)**
   - Extraer una muestra representativa de los datos.
2. **Exploración (Explore)**
   - Explorar los datos para encontrar patrones iniciales.
   - Visualizar los datos.

SEMMA: Fases Posteriores

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.

Metodología Ágil: Fases Iniciales

1. **Planificación**
   - Definir el alcance del proyecto y los objetivos principales.
2. **Desarrollo Iterativo**
   - Desarrollar el proyecto en ciclos cortos con incrementos frecuentes.

Metodología Ágil: Fases Posteriores

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

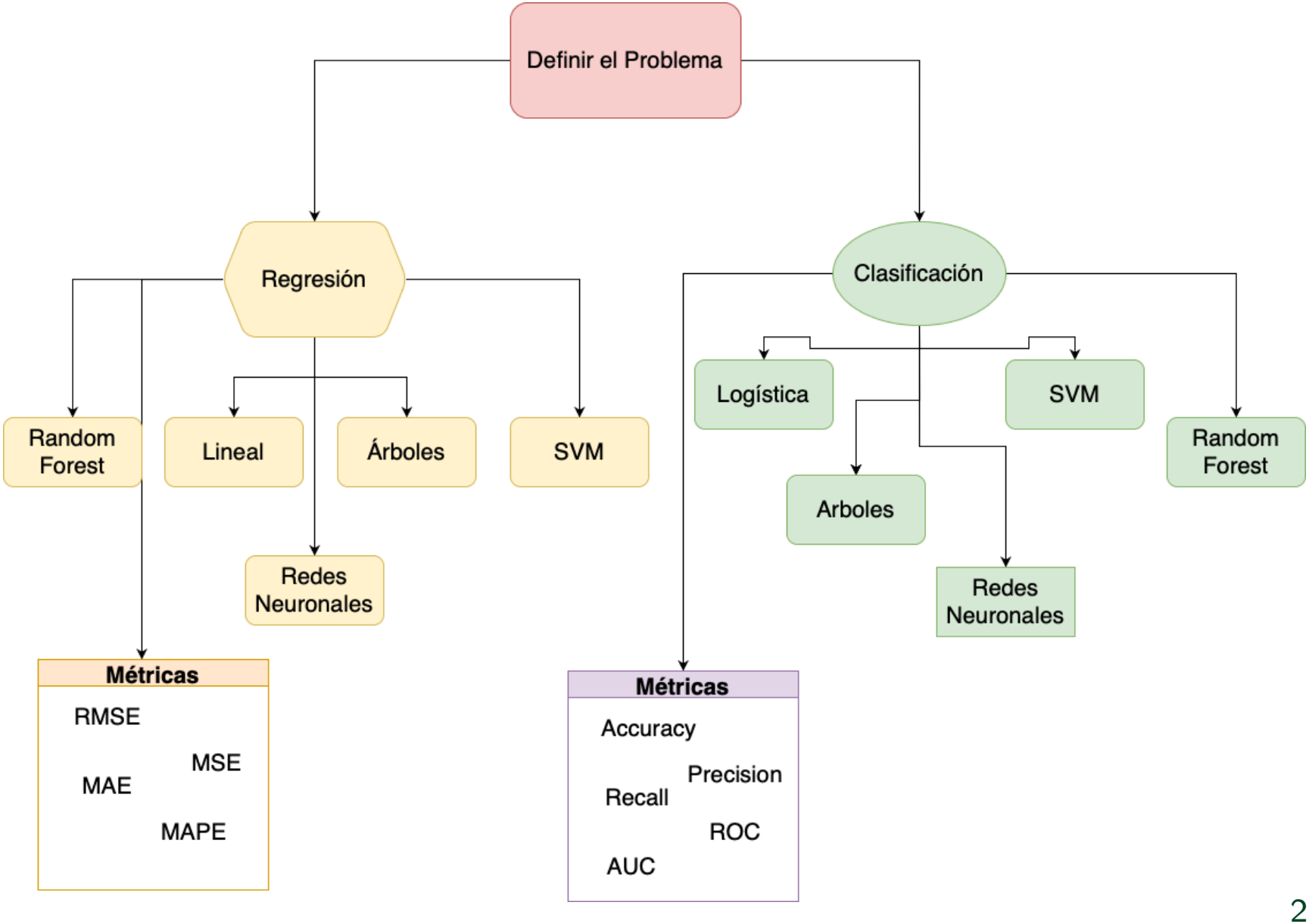
SEMMA

- 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



Inferencia

Figure 2: Hold out

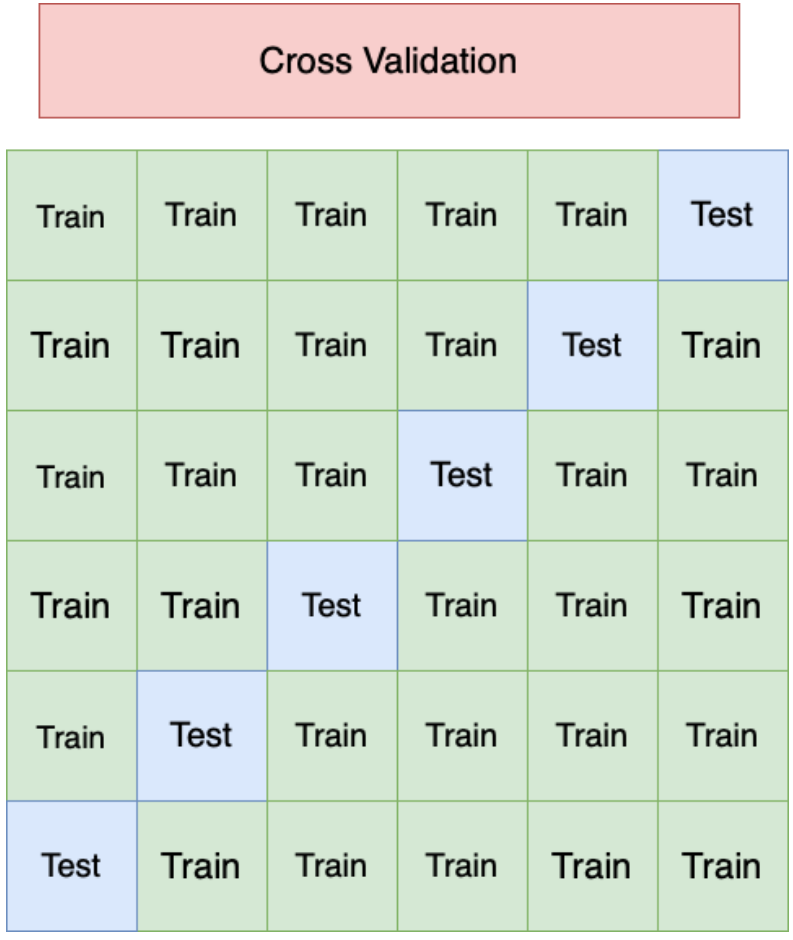
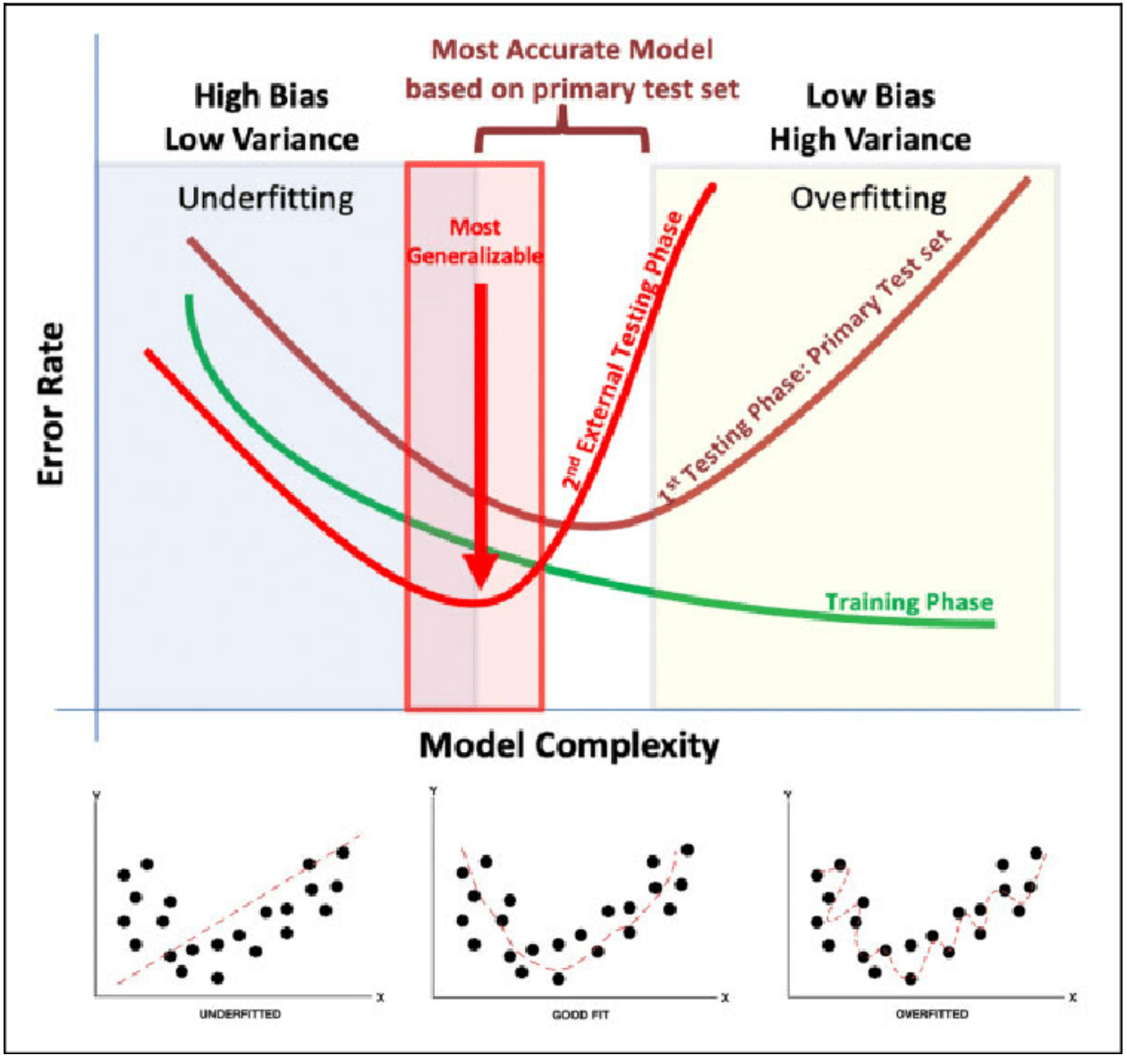


Figure 3: Cross Validation

Tradeoff


Most Accurate Model
based on primary test set
High Bias
Low Variance
Underfitting
Most
Generalizable
Low Bias
High Variance
Overfitting
2nd External Testing Phase
1st Testing Phase: Primary Test set
Training Phase
Error Rate
Model Complexity
UNDERFITTED
GOOD FIT
OVERFITTED