03 – Preparación de los Datos

Centro de Servicios y Gestión Empresarial
SENA Regional Antioquia

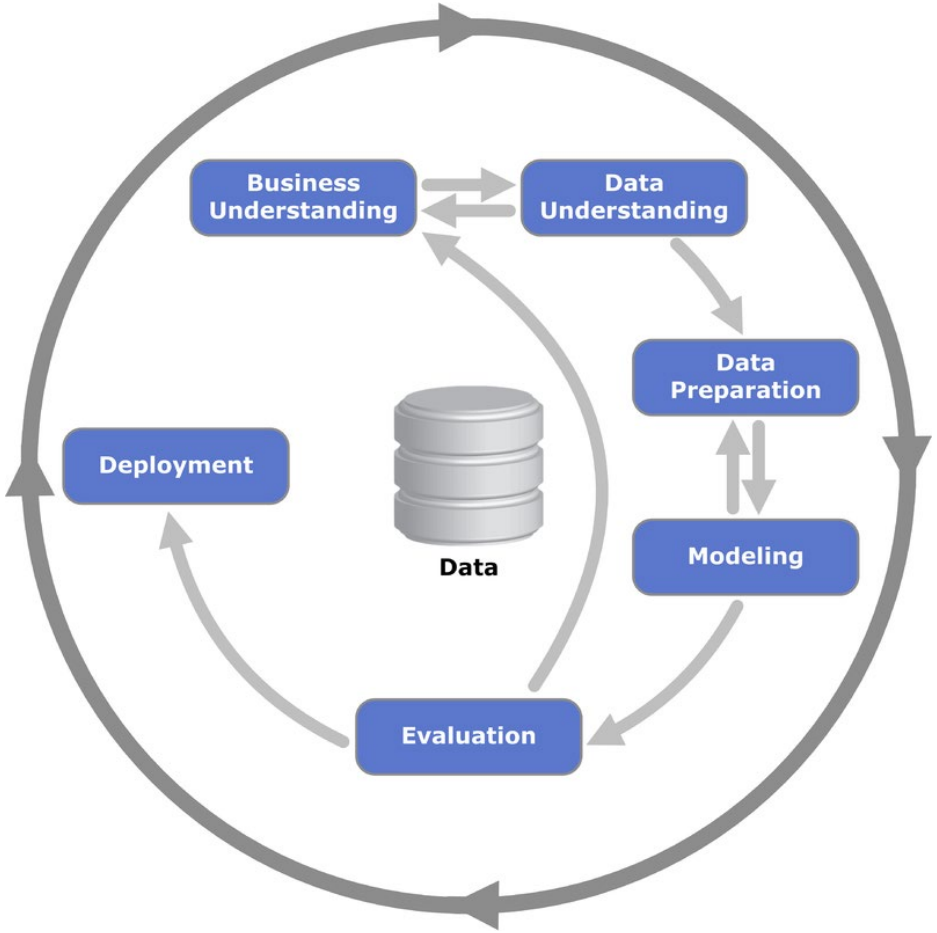
@SENAComunica
www.sena.edu.co

Preparación de los Datos

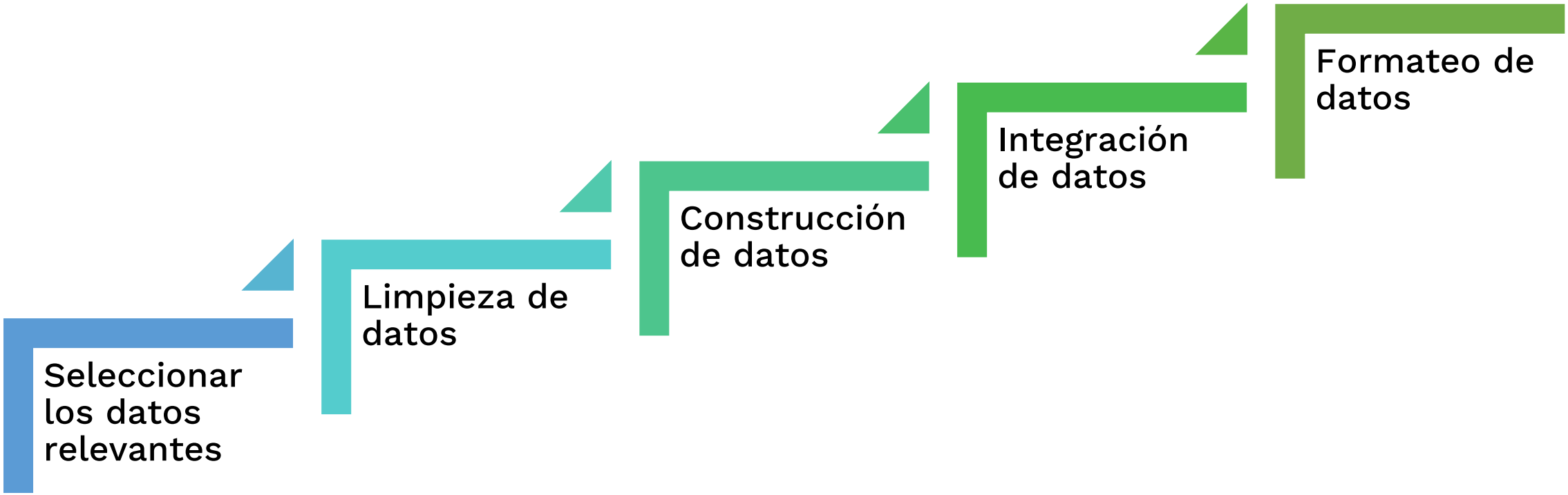
Alvaro Pérez Niño - Instructor

Preparación de los Datos

La fase de preparación de los datos en la metodología CRISP-DM; es crucial para el éxito de cualquier proyecto de ciencia de datos. En esta fase, **los datos crudos se transforman en un formato estándar que se puede usar para el modelado.**



Preparación de los Datos



Preparación de los Datos

1. **Seleccionar los datos relevantes:**

Decidir qué variables (columnas) y registros (filas) son relevantes para el análisis. A menudo, **no todos los datos disponibles son útiles**, por lo que se deben seleccionar aquellos que sean **más significativos** para el problema que se intenta resolver.

Se podría basar esta selección en el **conocimiento del negocio**, los resultados de **análisis previos**, o criterios específicos como la disponibilidad de datos o la relevancia para el problema a resolver.

Preparación de los Datos

2. Limpieza de datos:

- **Manejo de valores faltantes:** Identificar y manejar los datos ausentes. Esto puede implicar imputar valores (por ejemplo, con la media, mediana o un valor específico), eliminar filas o columnas con demasiados valores faltantes.

- **Detección y corrección de errores:** Revisar los datos para identificar y corregir errores, como entradas duplicadas, valores fuera de rango, o inconsistencias en datos categóricos.

Preparación de los Datos

3. Construcción de datos:

- **Generación de nuevas variables:** A veces es necesario crear nuevas variables que puedan ser más útiles para el análisis. Esto podría implicar calcular ratios, agrupar categorías, o aplicar transformaciones matemáticas a las variables existentes.

- **Transformación de datos:** Puede ser necesario transformar los datos para que se ajusten mejor a los requisitos del modelo. Esto podría incluir la normalización o estandarización de variables numéricas, o la conversión de variables categóricas en variables binarias/numericas.

Preparación de los Datos

4. Integración de datos:

- Si los datos provienen de diferentes fuentes, esta fase implica combinarlos en un solo dataset coherente.

- Esto puede incluir la unión de diferentes tablas o la fusión de datos históricos con datos actuales.

Preparación de los Datos

5. Formateo de datos:

- Adaptar el formato de los datos para que se ajusten a los requisitos de las técnicas de modelado que se van a utilizar.

- Esto podría incluir reordenar columnas, ajustar los tipos de datos, o asegurarse de que todas las variables están en el formato adecuado para ser procesadas.

G R A C I A S

Presentó: Alvaro Pérez Niño
Instructor Técnico
Correo: aperezn@sena.edu.co
http://centrodeserviciosygestionempresarial.blogspot.com/

Línea de atención al ciudadano: 01 8000 910270
Línea de atención al empresario: 01 8000 910682

@SENAComunica

www.sena.edu.co