03 – Preparación de los Datos

Centro de Servicios y Gestión Empresarial
SENA Regional Antioquia

@SENAComunica

www.sena.edu.co

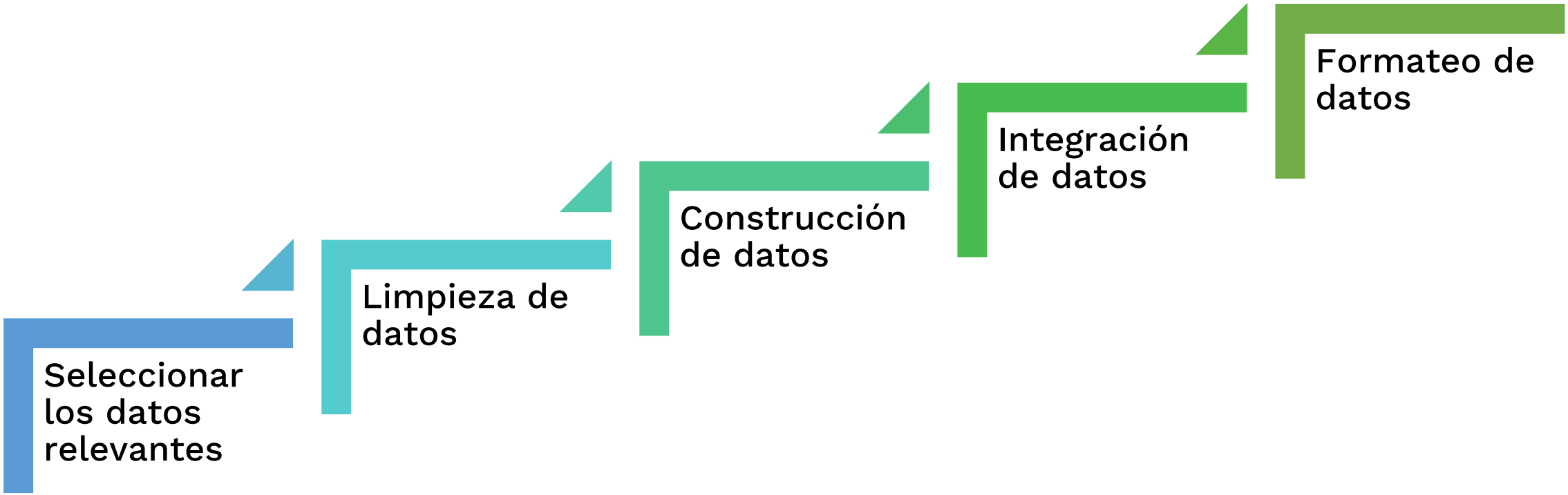
Preparación de los Datos

Preparación de los Datos

La fase de preparación de los datos en la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es crucial para el éxito de cualquier proyecto de ciencia de datos. En esta fase, **los datos crudos se transforman en un formato que se puede usar para el modelado.**

Variable	Tipo	Descripcion	Ejemplo
id	int	Identificador de la vivienda	1 2 3 4 5
zona	chr	Ubucación de la vivienda - Zona	“Zona Norte” “Zona Oriente”...
piso	int	Ubicacion horizontal de la vivienda	4 1 NA 2 NA NA 2 NA NA 2 ...
estrato	int	Clasificacion socioeconomica	6 6 5 3 5 5 6 5 5 5 ...
preciom	int	Valor económico en millones	64500000 43000000 450000000 ...
areaconst	num	Area construida de la vivienda	260 150 118 72 60 990 104 ...
parquea	int	Cantidad de parqueaderos disponibles	2 1 4 1 2 8 2 4 1 2 ...
banios	int	Cantidad de unidades sanitarias	2 1 4 1 2 8 2 4 1 2 ...
habitac	int	cantidad de habitaciones	2 1 4 1 2 8 2 4 1 2 ...
tipo	chr	Tipo de vivienda	“Casa” “Apartamento” ...
barrio	chr	Ubucación de la vivienda - barrio	“Normandia” “Brio. Granada” ...
latitud	num	Ubucación de la vivienda - latitud	-79571 -76568 -76565 -76565 ...
longitud	num	Ubucación de la vivienda - longitud	3454 3454 3455 3417 3408 ...

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.

Conceptos Estadísticos

Practicas de Python

Preparación de los Datos

Practicas de Python

LIMPIEZA Y TRANSFORMACIÓN DE DATOS

El proceso de preprocesamiento de datos se lleva a cabo una vez que hemos explorado y limpiado nuestro conjunto de datos, de modo que entendamos su contenido, estructura y calidad. Al explorar nuestros datos, es probable que obtengamos una buena idea de cómo queremos modelarlos, lo cual nos ayudará a decidir la mejor manera de preprocesarlos para que estén listos para e modelado desde el principio.

```
#Importar la librería Pandas e instalación de pandas-profiling
import pandas as pd
#!pip install pandas-profiling[notebook]
```

```
#Cargar el dataset
```

Ejercicio de Aplicación

03 – Preparación de los Datos

Análisis del mercado inmobiliario en Cali

La empresa B&C (Bines y Casas) es una agencia de bienes raíces que opera en la ciudad de Cali, Colombia. La empresa fue fundada por Sandra Milena hace 10 años y actualmente cuenta con ocho agentes de bienes raíces.

El mercado de bienes raíces en Cali ha crecido significativamente en los últimos años, impulsado por el crecimiento de la población, la inversión extranjera directa y el desarrollo de nuevos proyectos inmobiliarios. En 2022, las ventas del sector en Cali llegaron a $6700 millones y en 2023 a $6100 mil millones. Se espera que este sector continue creciendo durante los próximos años, permitiendo un desarrollo dinámico en la economía regional.

Análisis del mercado inmobiliario en Cali

La empresa B&C cuenta con información sobre viviendas que incluye información sobre el precio, la ubicación, las características y la venta de viviendas en Cali.

Esta información gestionada mediante ciencia de datos, sería útil para la empresa B&C en la tomar decisiones sobre su negocio, tales como:

- Definir su nicho de mercado.
- Desarrollar estrategias de marketing.
- Establecer precios de venta.
- Ofrecer servicios personalizados a sus clientes.

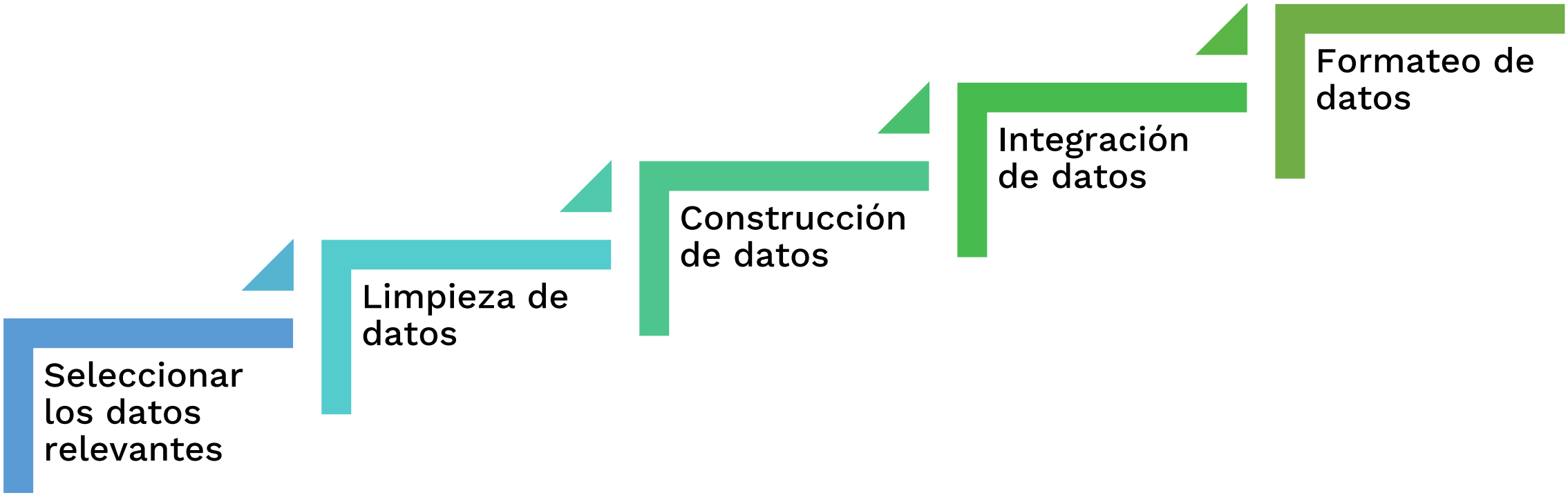
Dataset_Mobiliario.csv

Archivo Editar Ver

```
id,zona,piso,estrato,preciom,areaconst,parquea,banios,habitac,tipo,barrio,longitud,latitud
8312,Zona Oeste,4,6,1300,318,2,4,2,Apartamento,arboleda,-76576,3454
8311,Zona Oeste,1,6,480,300,1,4,4,Casa,normandía,-76571,3454
8307,Zona Oeste,NA,5,1200,800,4,7,5,Casa,miraflores,-76568,3455
8296,Zona Sur,2,3,220,150,1,2,4,Casa,el guabal,-76565,3417
8297,Zona Oeste,NA,5,330,112,2,4,3,Casa,bella suiza alta,-76565,3408
8298,Zona Sur,NA,5,1350,390,8,10,10,Casa,bella suiza alta,-76565,3409
8299,Zona Sur,2,6,305,125,2,3,3,Apartamento,bella suiza,-76565,3408
8300,Zona Oeste,NA,5,480,280,4,4,4,Apartamento,bella suiza alta,-76565,3408
8286,Zona Sur,NA,5,275,74,1,2,3,Apartamento,valle del lili,-76564,3409
8287,Zona Sur,2,5,285,120,2,4,3,Apartamento,bella suiza,-76564,3.41
8288,Zona Sur,1,5,310,166,2,4,3,Apartamento,bella suiza,-76564,3.41
8281,Zona Oeste,NA,3,175,155,NA,4,6,Casa,el nacional,-76563,3421
8282,Zona Oeste,6,6,640,157,2,3,3,Apartamento,santa teresita,-76563,3418
8283,Zona Oeste,NA,3,98,60,NA,2,3,Apartamento,aguacatal,-76563,3458
8274,Zona Oeste,2,5,416,98,1,2,2,Apartamento,santa teresita,-76562,3418
8275,Zona Oeste,8,6,700,123,2,3,4,Apartamento,bellavista,-76562,3423
8276,Zona Oeste,NA,5,393,131,2,3,3,Apartamento,bellavista,-76562,3423
8277,Zona Oeste,6,6,700,240,2,5,4,Apartamento,el peñon,-76562,3.45139
8265,Zona Sur,NA,6,1390,350,4,7,3,Casa,la riverita,-76561,3455
8266,Zona Oeste,NA,6,360,90,1,1,2,Apartamento,normandía,-76561,3453
8267,Zona Oeste,3,3,110,55,NA,1,3,Apartamento,miradol del aguacatal,-76561,3461
8268,Zona Oeste,NA,3,142,50,NA,1,2,Apartamento,aguacatal,-76561,3455
8269,Zona Oeste,NA,4,123,60,NA,1,2,Apartamento,aguacatal,-76561,3455
8226,Zona Oeste,NA,3,165,114,NA,4,6,Casa,terrón colorado,-76559,3452
8227,Zona Oeste,1,6,1400,300,2,4,3,Apartamento,normandía,-76559,3375
8228,Zona Sur,1,3,145,60,1,2,3,Apartamento,mel√©ndez,-76559,3386
8229,Zona Sur,8,4,240,90,1,2,3,Apartamento,guadalupe,-76559,3417
8230,Zona Oeste,4,6,440,97,2,2,2,Apartamento,cristales,-76559,3419
8231,Zona Oeste,9,6,750,150,2,4,3,Apartamento,cristales,-76559,3419
8232,Zona Oeste,2,6,900,250,NA,5,4,Apartamento,bellavista,-76559,3419
8233,Zona Oeste,NA,6,900,342,3,5,3,Apartamento,la arboleda,-76559,3388
8234,Zona Oeste,NA,6,1100,320,3,4,3,Apartamento,aguacatal,-76559,3458
```

Dataset

Preparación de los Datos



LMS

ybkyfpp

Ciencia de Datos - 2742550 Tg. ADSO Copiar enlace de invitación

G R A C I A S

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