

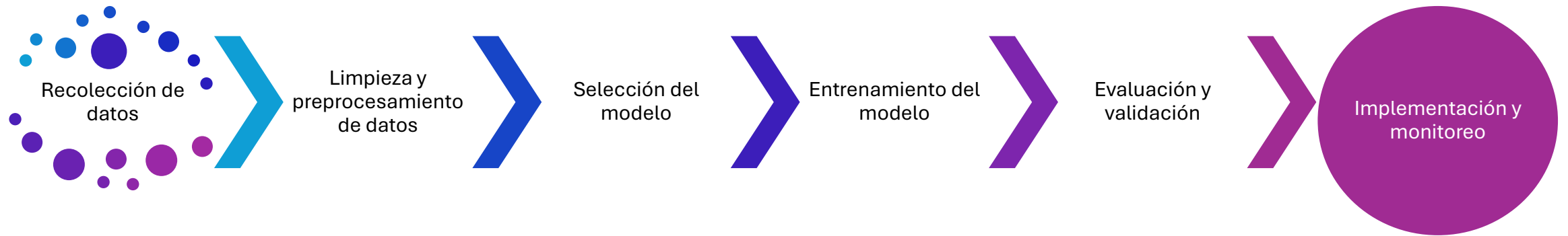
Machine Learning



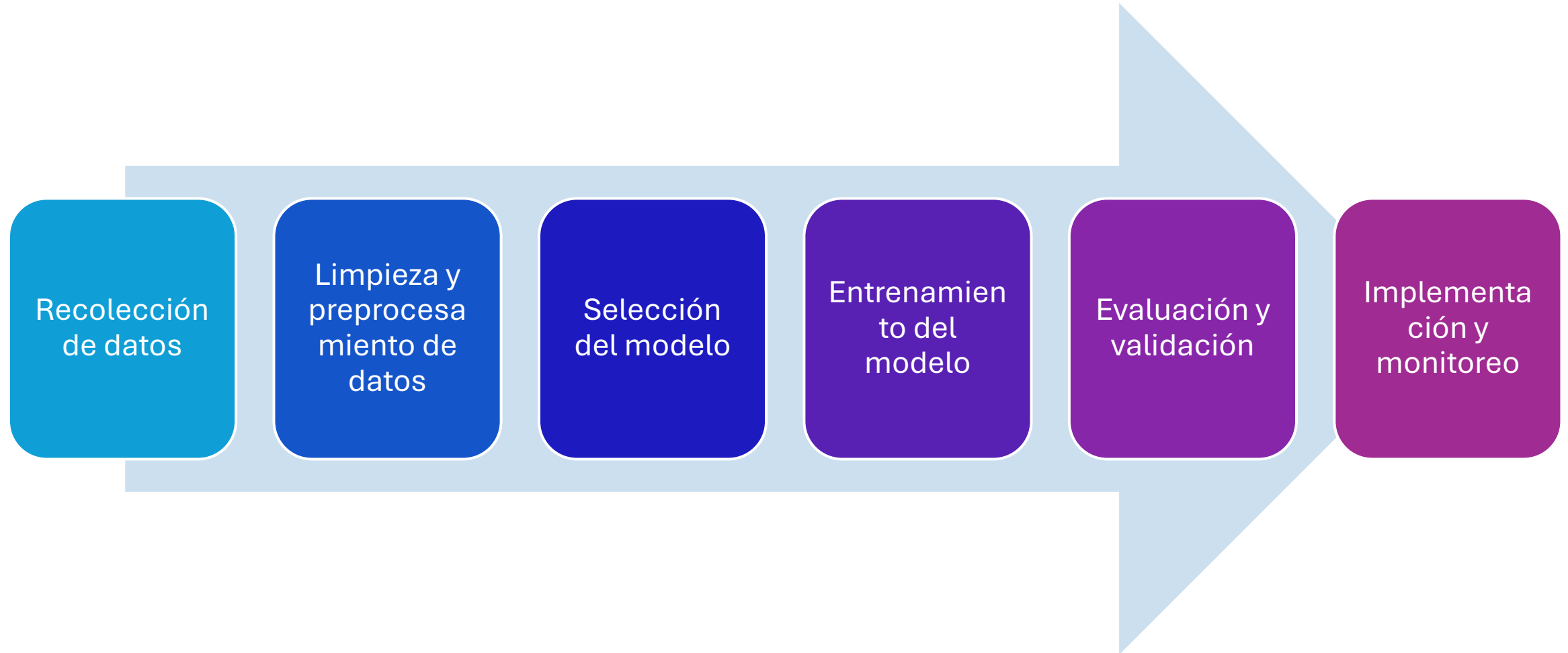
Susana Medina Gordillo

susana.medina@correounivalle.edu.co

Flujo de trabajo en Machine Learning



Flujo de trabajo en Machine Learning



Preprocesamiento de Datos

Preprocesamiento de Datos: Introducción

- Limpieza de datos y manejo de datos faltantes
- Normalización y estandarización

¿Por qué es crucial el preprocesamiento?

La **calidad** de los datos es fundamental para el **éxito** de cualquier modelo de Machine Learning.

Los datos del mundo real suelen ser "sucios": incompletos, inconsistentes, ruidosos y en formatos no ideales.

El preprocesamiento transforma los datos brutos en un **formato limpio** y adecuado para el modelado.

Limpieza de datos

Limpieza de Datos

- Es una **tarea** básica de la Ciencia de Datos.
- Es fundamental preparar los datos ANTES del análisis.
- Algunas tareas de la limpieza de datos:
 - Encontrar / Remover **duplicados**
 - Encontrar / Remover datos **incompletos**
 - Encontrar / Remover datos **anómalos** (*outliers*)
 - **Modificar** y **validar** datos
 - **Rectificar** registros dudosos

Limpieza de Datos

Manejo de valores faltantes:

- Eliminación de filas o columnas con muchos valores faltantes.
- Imputación: rellenar los valores faltantes con la media, mediana, moda u otros métodos más sofisticados.

Eliminación de duplicados:

- Identificar y eliminar registros duplicados para evitar sesgos en el modelo.

Corrección de errores:



- Identificar y corregir valores incorrectos o inconsistentes (ej., errores de tipeo, formatos incorrectos).

Manejo de valores atípicos (*outliers*):

- Identificar y tratar valores extremos que pueden distorsionar el modelo.

**Y si nos
saltamos la
limpieza de
datos?**





 + Machine Learning = 



Data

 + Artificial Intellifence = 

Data

 + Generative AI = 

Data

 + Agentic AI = 

Data

Normalización y estandarización

Normalización y Estandarización: ¿Por qué son necesarias ?

- ✓ Muchos algoritmos de *Machine Learning* son **sensibles** a la **escala de las variables**.
- ✓ La normalización y estandarización aseguran que **todas las variables tengan un rango similar**, lo que mejora el rendimiento del modelo.

Normalización y Estandarización: Técnicas

- ✓ **Normalización** (*Min-Max Scaling*): Escala los valores entre **0 y 1**.
- ✓ **Estandarización** (*Z-Score*): Transforma los valores para que tengan:
 - ✓ Media = 0
 - ✓ Desviación estándar = 1

Análisis Exploratorio de Datos (EDA)

- Visualización de datos con Matplotlib y Seaborn.
- Detección de patrones y relaciones en los datos.

Visualización de Datos: librerías python

➤ **Matplotlib:** Librería básica para crear gráficos estáticos, interactivos y de animación en Python.

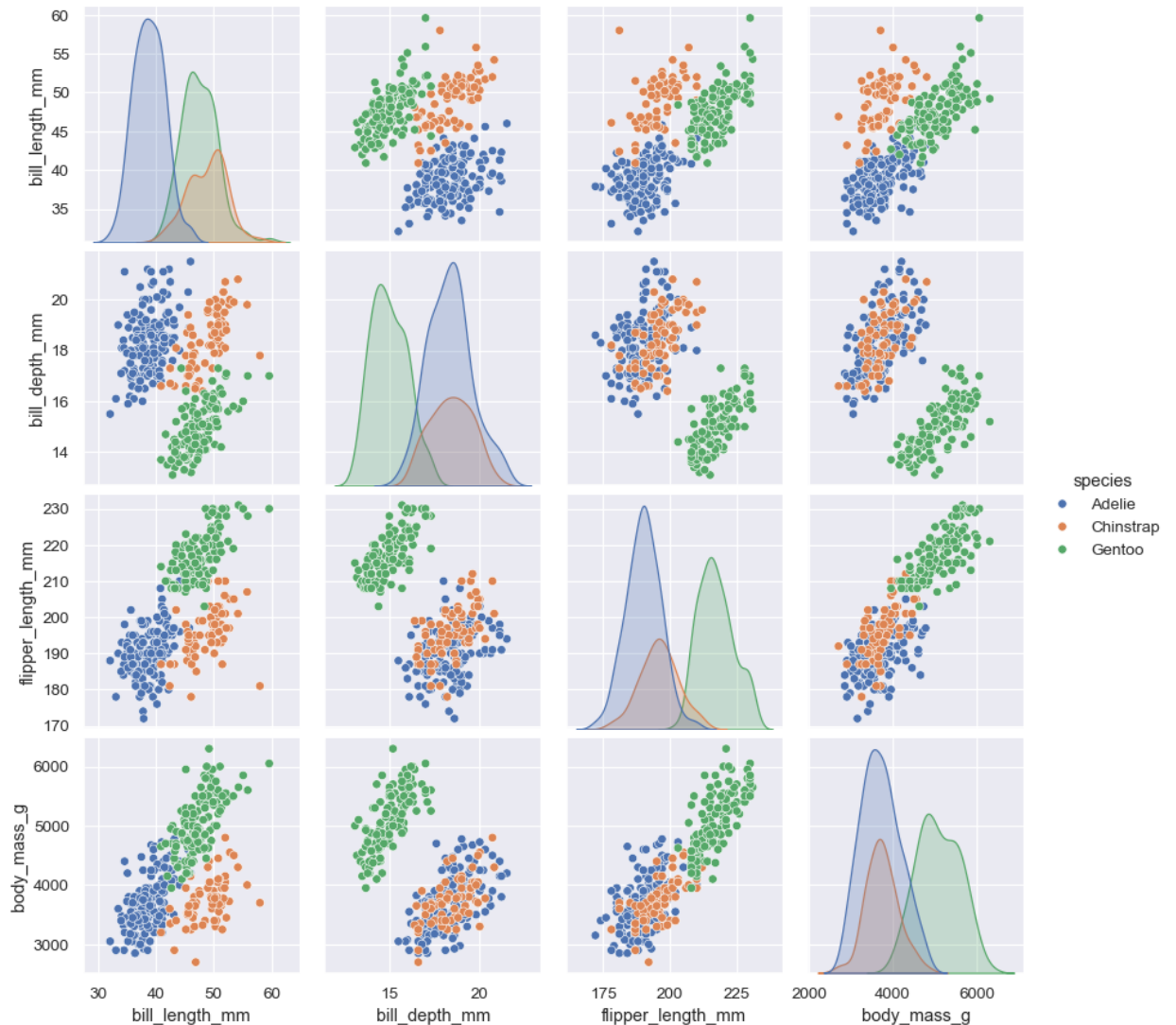


➤ **Seaborn:** Construida sobre Matplotlib, proporciona una interfaz de alto nivel para crear gráficos estadísticos informativos y atractivos.



Visualización de Datos: tipos de gráficos

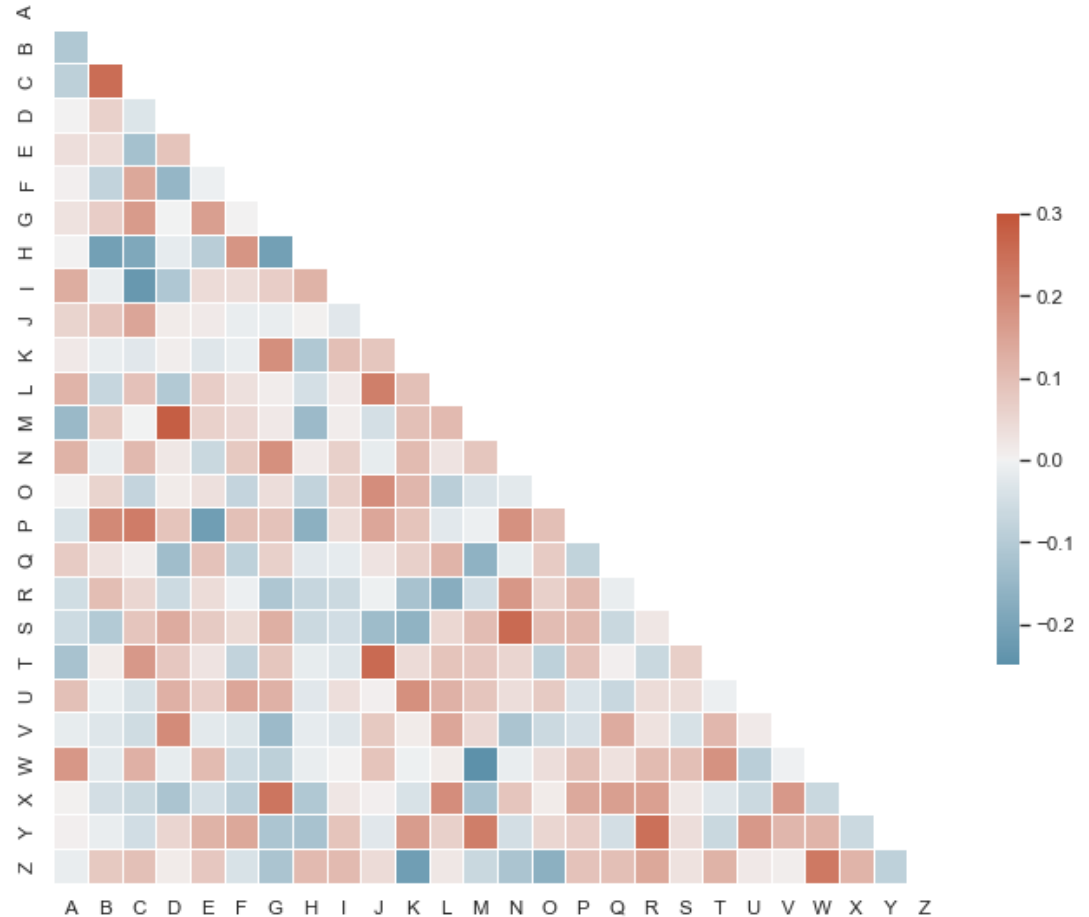
- Histogramas
- Diagramas de dispersión
- Gráficos de barras
- Boxplots / Diagramas de cajas



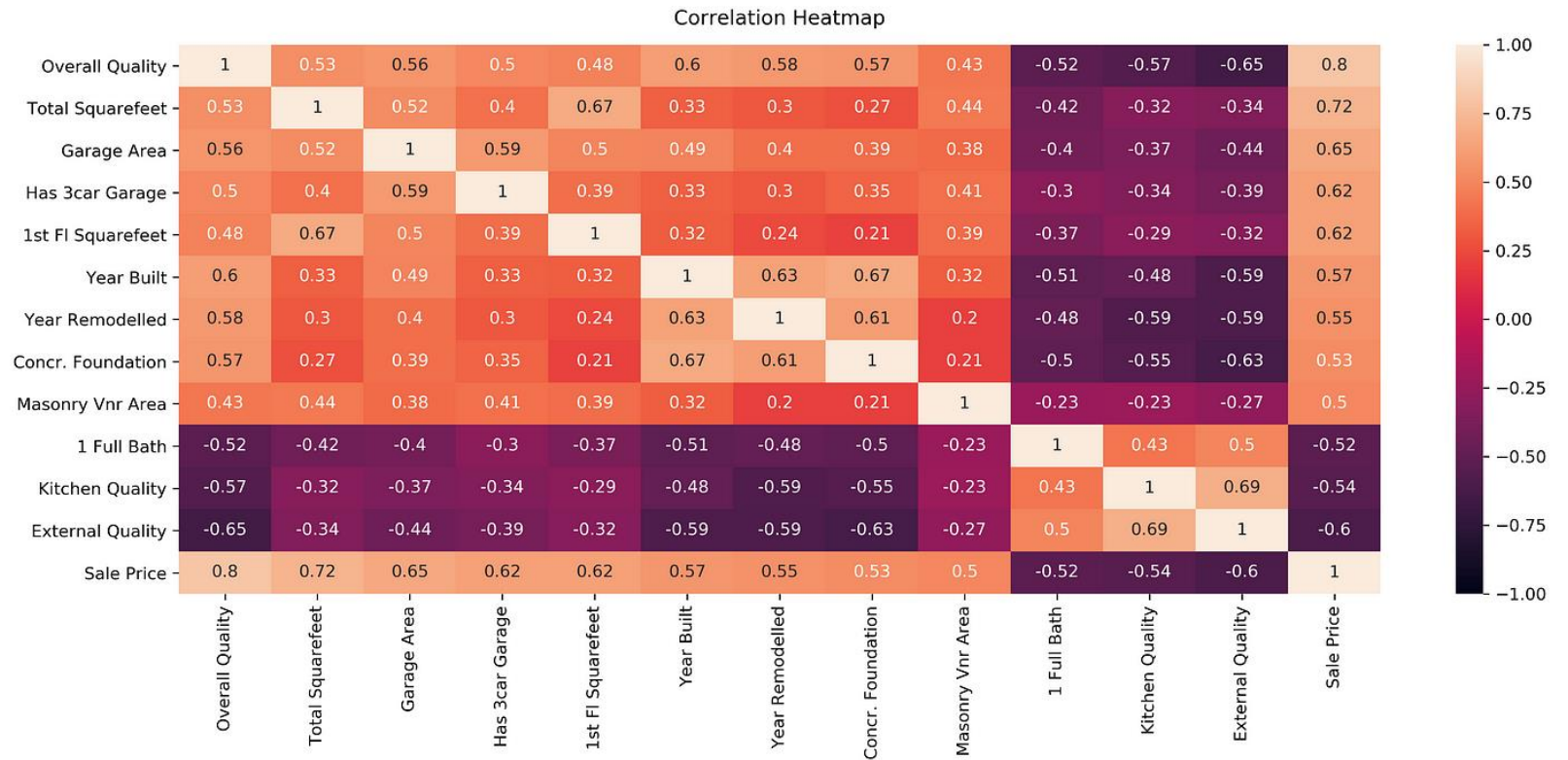
Detección de Patrones y Relaciones

¿Qué buscamos?

- **Tendencias:** ¿Los datos aumentan, disminuyen o se mantienen constantes con el tiempo?
- **Correlaciones:** ¿Existe una relación entre dos o más variables?
- **Patrones:** ¿Hay grupos o *clusters* en los datos?



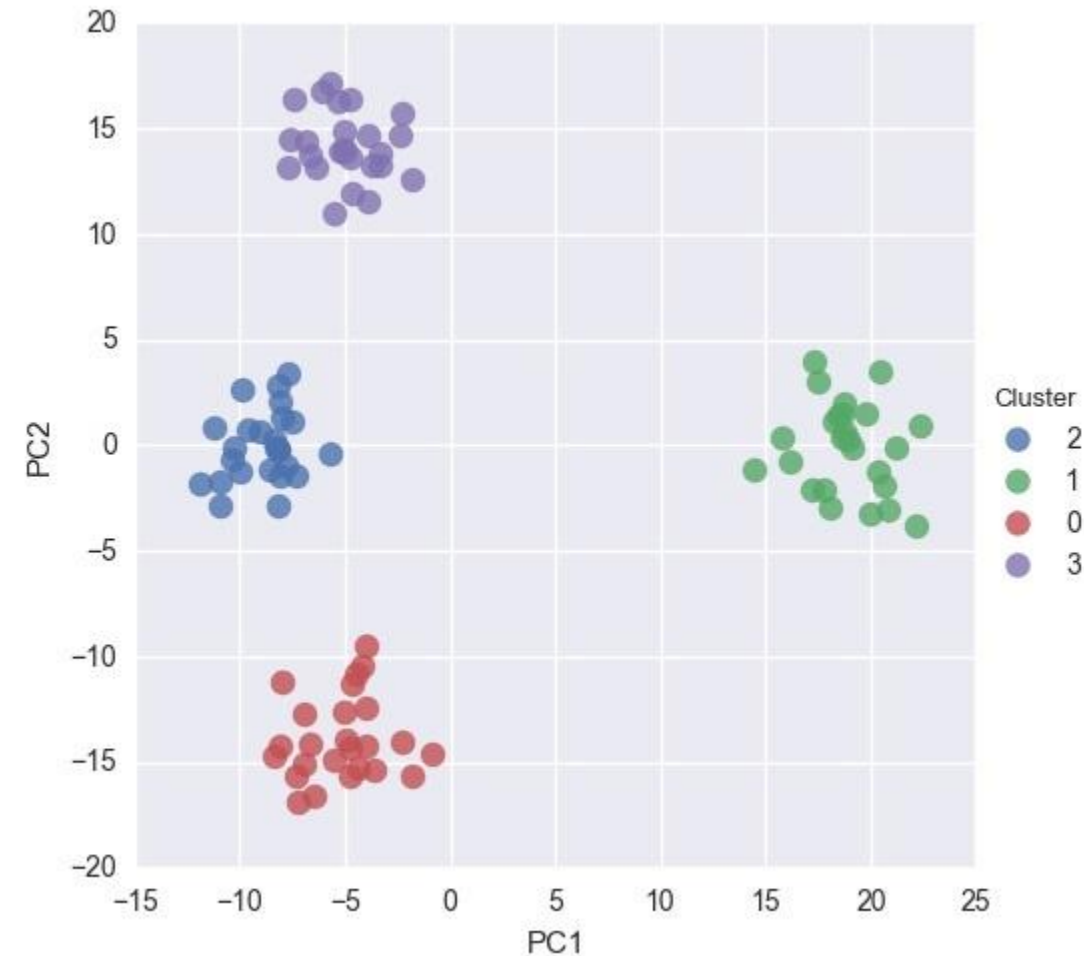
Detección de Patrones y Relaciones: Correlación (heatmap)



Detección de Patrones y Relaciones

¿Cómo lo hacemos?

- Visualización de datos
- Estadísticas descriptivas
- Análisis de componentes principales (**PCA**)



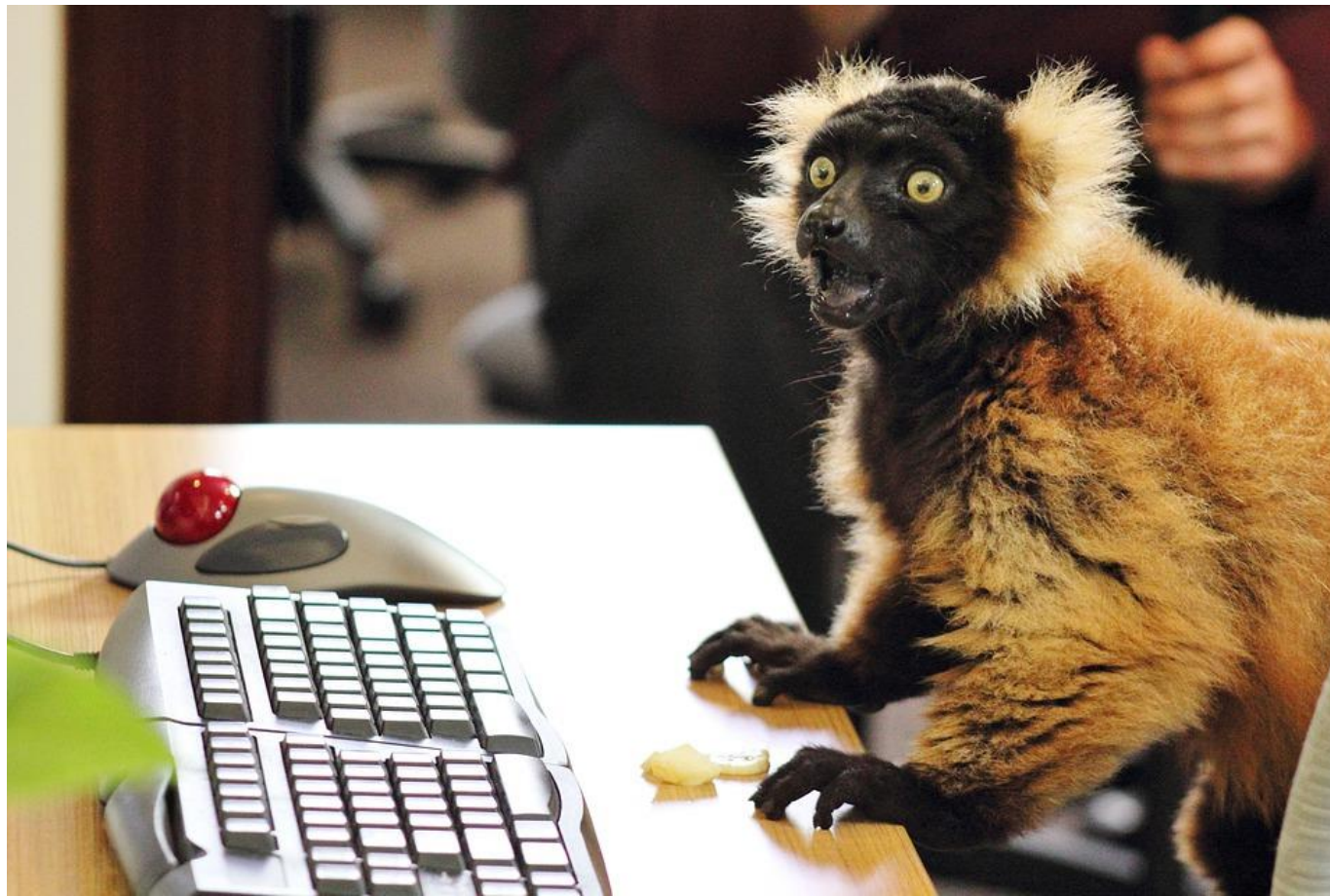
Conclusiones...

El preprocesamiento de datos es una **etapa esencial** en cualquier proyecto de Machine Learning.

La limpieza de datos, el manejo de valores faltantes, la normalización, la estandarización y el EDA son **técnicas fundamentales** que mejoran la calidad de los datos y el **rendimiento** del modelo.

Ejercicio práctico

colab.research.google.com



Ejercicio práctico: Google Colaboratory (*Colabs*)

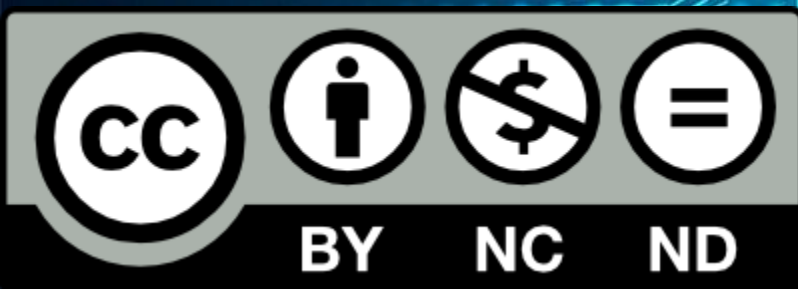
- Página oficial: <https://colab.google/>
- Abrir Colab (incluye tutorial): <https://colab.research.google.com/>
- Guía para EDA: https://colab.research.google.com/github/Tanu-N-Prabhu/Python/blob/master/Exploratory_data_Analysis.ipynb
- Video, Introducción a Google Colab:
<https://www.youtube.com/watch?v=9g61bnipcSs>



Referencias

- “Data Cleaning: Understanding the Essentials”. Consultado: el 11 de febrero de 2025. [En línea]. Disponible en: <https://www.datacamp.com/tutorial/tutorial-data-cleaning-tutorial>
- “Python Seaborn Tutorial For Beginners: Start Visualizing Data”. Consultado: el 10 de febrero de 2025. [En línea]. Disponible en: <https://www.datacamp.com/tutorial/seaborn-python-tutorial>
- “Python Boxplots: A Comprehensive Guide for Beginners”. Consultado: el 6 de febrero de 2025. [En línea]. Disponible en: <https://www.datacamp.com/tutorial/python-boxplots>
- Información e ideas presentadas basadas en el conocimiento general de modelos de lenguaje de IA. Gemini 2.9 Flash. Consultado: el 10 de febrero de 2025. [En línea].
- Imagen “Poop data” de Dr. Christian Krug. Perfil de [LinkedIn](#). Publicación de Eduardo Ordax: <https://www.linkedin.com/feed/update/urn:li:activity:7275487825238634496/>
- Imágenes de seaborn plots:
 - https://seaborn.pydata.org/examples/many_pairwise_correlations.html
 - https://seaborn.pydata.org/examples/scatterplot_matrix.html
 - <https://cmdlinetips.com/2018/03/pca-example-in-python-with-scikit-learn/>

Machine Learning



Susana Medina Gordillo

susana.medina@correounivalle.edu.co