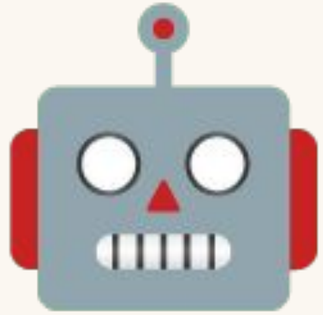
Corporación Universitaria Comfacauca
Unicomfacauca

ACREDITACIÓN INSTITUCIONAL EN ALTA CALIDAD
Resolución 12513 del 29/06/22. Sede principal.
Vigilada MinEducación

Desbloqueando el Poder de los Datos



Inteligencia Artificial & Ciencia de Datos para todos

Comenzamos a las 7:05 a.m. en punto.

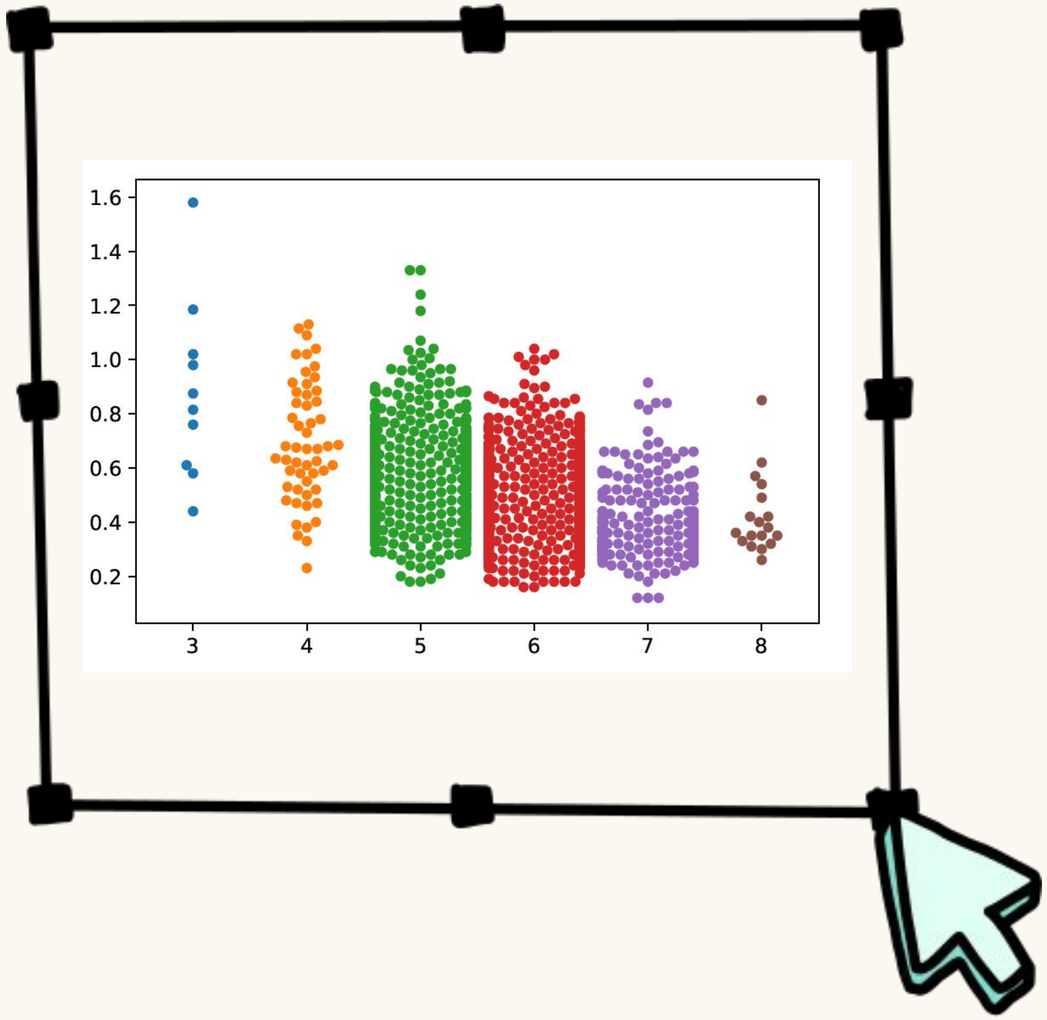
¿Te gustaría comenzar el día con alguna canción en específico?

Coméntala en el chat 🎶💬

(Opcional)

Encuesta anónima: ¿Cómo va el curso?

https://forms.gle/dYPNuZosMmgvjKUU9



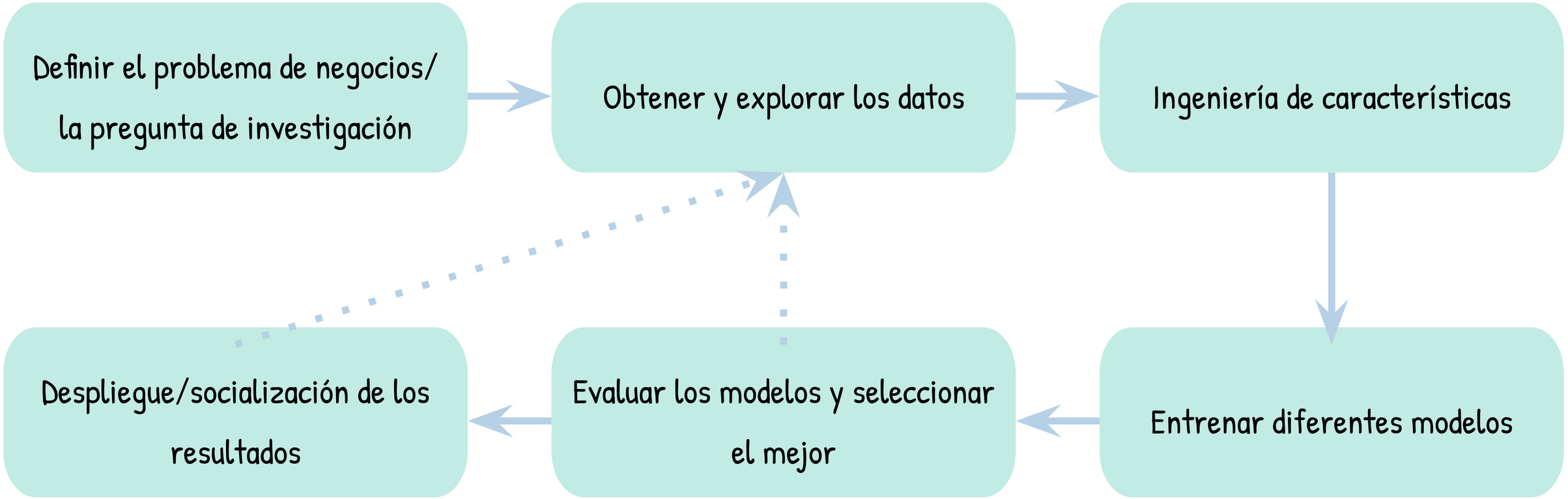
Exploración de datos

Septiembre 17, 2024

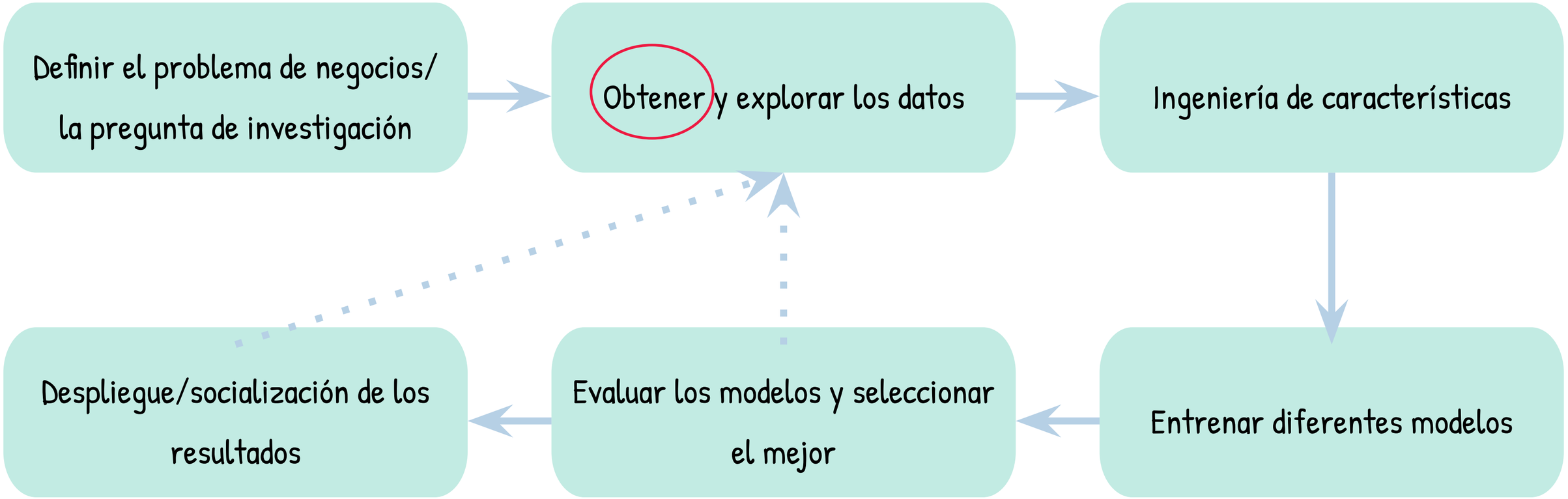
Agenda

1. Repaso de la última clase
2. Tema de hoy:
   - Exploración de datos

Pasos en un proyecto de Machine Learning



Pasos en un proyecto de Machine Learning

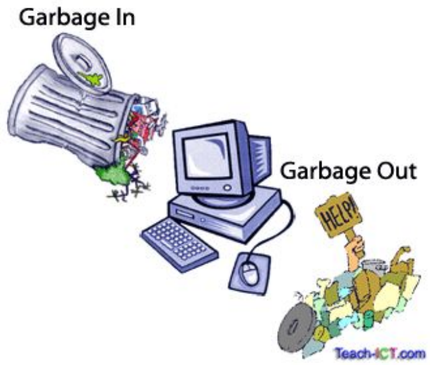


Obtener datos para un proyecto de Machine Learning

- La adquisición de datos es el proceso de identificar, recopilar y extraer información útil de diversas fuentes para su uso en proyectos de ciencia de datos y aprendizaje automático

- Dado que los datos se han convertido en un recurso tan valioso como el petróleo en la economía digital, una adecuada adquisición garantiza que los modelos de aprendizaje automático tengan una base sólida para su entrenamiento.

- La calidad, relevancia y variedad de los datos obtenidos influyen directamente en la efectividad, precisión y desempeño del modelo, haciendo que los datos sean un componente esencial para el éxito del proyecto.



Datos estructurados vs datos no estructurados

Ventas

Producto	Potencia	Unidades	Ganancias
Bicicletas	Eléctrica	476	$751.604
Bicicletas	Manual	302	$581.350
Motonetas	Eléctrica	387	$427.248
Motonetas	Manual	309	$48.513
Patinetas	Eléctrica	251	$135.791
Bicicletas	Eléctrica	354	$558.966
Bicicletas	Manual	219	$336.165
Motonetas	Eléctrica	312	$583.128
Motonetas	Manual	419	$396.793

- Los **datos estructurados** están altamente organizados y son fácilmente legibles por máquinas. Normalmente se almacenan en formatos tabulares, como hojas de cálculo (CSV, Excel) o bases de datos relacionales (SQL).

  Cada observación está en un fila y sus características en columnas predefinidas, lo que facilita su procesamiento y análisis.

- Los **datos no estructurados** no siguen un formato o estructura específica, lo que los hace más difíciles de organizar y analizar. Este tipo de datos incluye texto libre, imágenes, videos, audios y otros formatos multimedia.

  Debido a su naturaleza, los datos no estructurados a menudo requieren técnicas avanzadas, como procesamiento de lenguaje natural (NLP) o redes neuronales convolucionales (CNN).

¿Cómo conseguir datos?

1. **Tus propios datos**

   - Datos que generas o recopilas tú mismo, como encuestas, formularios, experimentos o datos de tu empresa/universidad.

   - Puedes cargarlos desde archivos locales (CSV, Excel, SQL, JSON) o desde la nube (Google Drive, AWS S3, etc.).

2. **Datos de código abierto**

   - Los conjuntos de datos de código abierto son colecciones de datos disponibles de manera gratuita, que cualquier persona puede usar, modificar y compartir.

   - Universidades, gobiernos y organizaciones de investigación también publican a menudo conjuntos de datos abiertos.

¿Cómo conseguir datos?

3. **APIs**

   - Una API (Interfaz de Programación de Aplicaciones) es una interfaz que permite acceder y recopilar datos de diversas fuentes de manera automatizada, facilitando la obtención de grandes volúmenes de información en ciencia de datos para su análisis.

   - Ejemplos incluyen la API de Twitter, Google Maps, Spotify, o APIs financieras para datos de mercado.

4. **Web Scraping**

   - Extraer datos de sitios web que no tienen una API disponible, pero permiten el acceso público a sus datos. Herramientas como BeautifulSoup, Scrapy, o Selenium te permiten automatizar este proceso.

¿Cómo conseguir datos?

5. **Bases de datos**
    - **SQL:** Obtener datos de bases de datos relacionales como MySQL, PostgreSQL, o SQLite.
    - **NoSQL:** Obtener datos de bases de datos NoSQL como MongoDB o Firebase.
6. **Comprar datos**
    - Plataformas donde puedes comprar o descargar conjuntos de datos, como Quandl o AWS Data Exchange.
7. **Simulación de datos**
    - Si no tienes acceso a datos reales, puedes generar datos sintéticos o simulados usando herramientas como scikit-learn o Faker.

Datos abiertos

1. **Repositorios de datos abiertos**
   - OpenML.org https://openml.org
   - Kaggle.com https://kaggle.com/datasets
   - PapersWithCode.com https://paperswithcode.com/datasets
   - UC Irvine Machine Learning Repository https://archive.ics.uci.edu/ml
   - Amazon's AWS datasets https://registry.opendata.aws
   - TensorFlow datasets https://tensorflow.org/datasets
   - Google's data search engine https://datasetsearch.research.google.com/

2. **Portales que tienen un listado de datos**
   - DataPortals.org https://dataportals.org/
   - Listado de Wikipedia https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research
   - Quora's list https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public
   - Reddit's dataset https://www.reddit.com/r/datasets
   - GitHub * https://github.com/

Datos abiertos

3. **Específicos por lugar**
   - San Francisco Open Data https://datasf.org/opendata/
   - NYC Open Data https://opendata.cityofnewyork.us/
   - Datos Abiertos Londres https://opendata.london.ca/
   - Datos Abiertos Colombia https://www.datos.gov.co/
   - Datos Abiertos Bogotá https://datosabiertos.bogota.gov.co/about
   - Burundi Open Data for Africa: https://burundi.opendataforafrica.org/
   - Datos Abiertos Popayán https://www.popayan.gov.co/Ciudadanos/Paginas/Datos-Abiertos-Alcaldia-de-Popayan.aspx#gsc.tab=0
   - Datos Abiertos Cauca https://www.cauca.gov.co/NuestraGestion/Paginas/Datos-Abiertos.aspx

Web Scraping con Beautiful Soup

- Beautiful Soup es una librería de Python que se utiliza para raspar y analizar documentos HTML o XML, facilitando la extracción de datos específicos de sitios web.

Una pequeña introducción a HTML

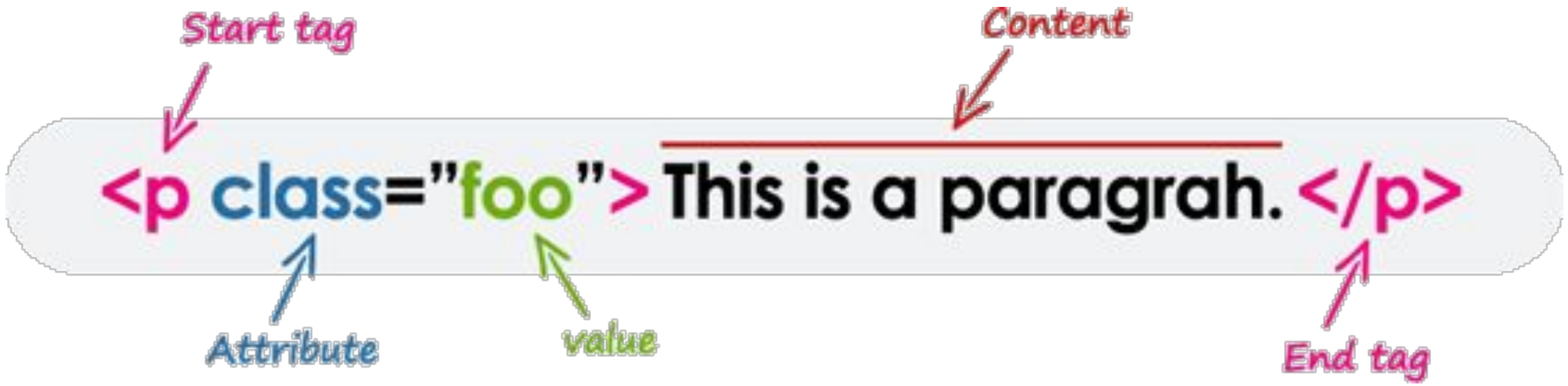
- El HTML es la estructura de una página web. Utiliza etiquetas para definir elementos, por ejemplo:
    - <h1>This is a heading. Esto es un título.</h1>
    - This is a paragraph. Esto es un párrafo.
    - This is a link. Esto es un enlace.

- **Etiquetas:** Elementos como <h1>, , <a>
- **Atríbutos:** Propiedades como *href*
- **Estructura de árbol:** El HTML forma una estructura de árbol, y así es que Beautiful Soup navega el contenido.

Una pequeña introducción a HTML

- El HTML es la estructura de una página web. Utiliza etiquetas para definir elementos, por ejemplo:
    - <h1>This is a heading. Esto es un título.</h1>
    - This is a paragraph. Esto es un párrafo.
    - This is a link. Esto es un enlace.

- **Etiquetas:** Elementos como <h1>, , <a>
- **Atríbutos:** Propiedades como *href*
- **Estructura de árbol:** El HTML forma una estructura de árbol, y así es que Beautiful Soup navega el contenido.


Start tag
Content
This is a paragrah.
Attribute
value
End tag

Ejemplo:

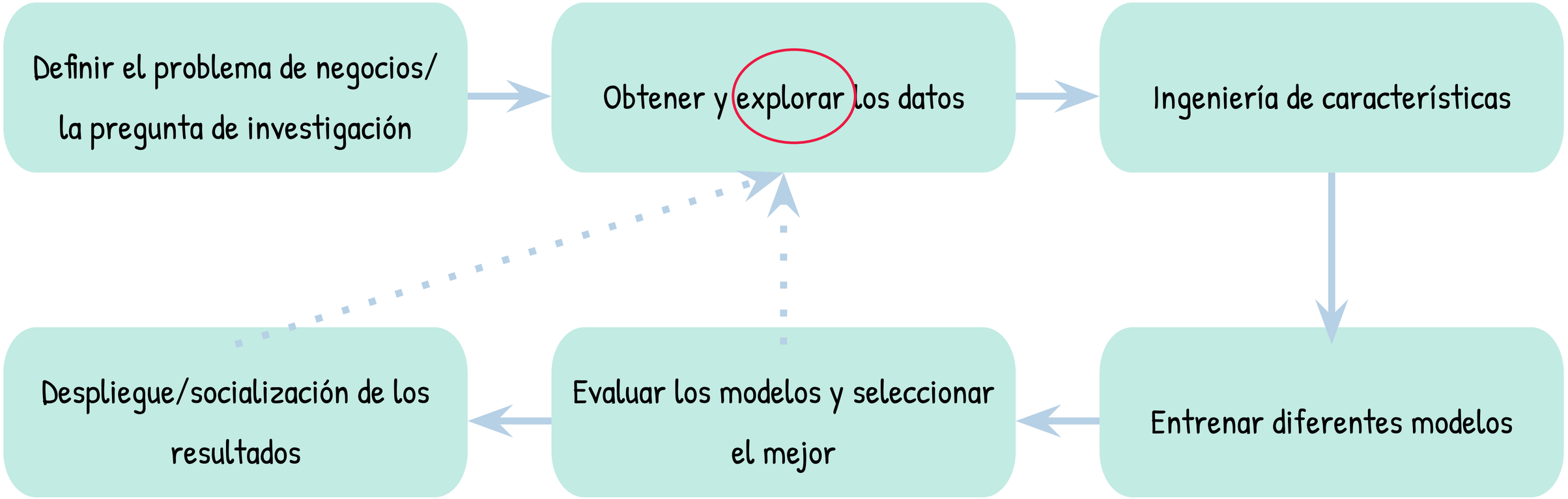
```
<a href="http://www.google.com/" id="buscador">Google</a>
```

1. ¿Cuál es la etiqueta?
2. ¿Cuáles son los atributos?
3. ¿Cuáles son los valores del atributo?
4. ¿Cuál es el contenido de la etiqueta?

Notebook de hoy

https://colab.research.google.com/drive/1PJw_z5QhoimnpZk5Pqd5VODcoSjemW01?usp=sharing

Pasos en un proyecto de Machine Learning



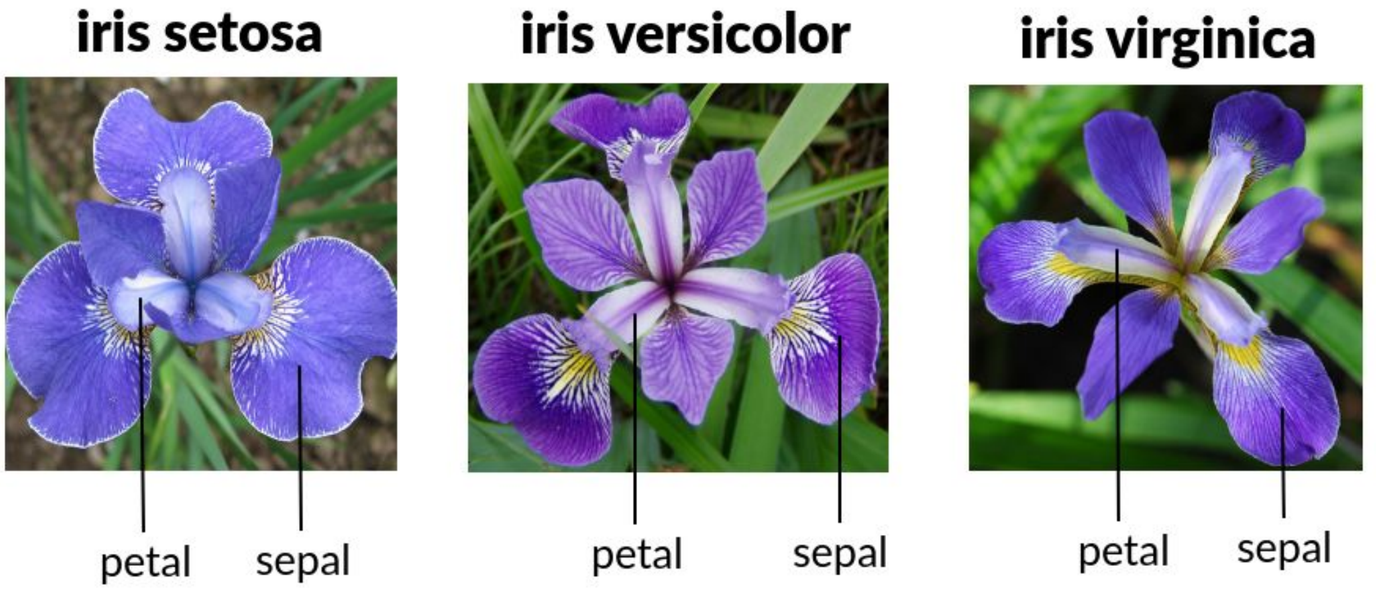
Análisis Exploratorio de Datos (EDA)

- El Análisis Exploratorio de Datos (EDA) es un proceso inicial en el análisis de datos, cuyo objetivo principal es obtener una comprensión profunda del conjunto de datos. Se utiliza para explorar, resumir y visualizar los datos, antes de aplicar cualquier modelo de machine learning.

Importancia del EDA en machine learning:

- **Identificación de patrones y relaciones:** Permite descubrir patrones importantes en los datos que podrían influir en los resultados del modelo.
- **Detección de errores y anomalías:** El EDA ayuda a identificar valores atípicos, errores de medición o datos faltantes que pueden afectar negativamente el rendimiento de los modelos.
- **Comprensión del contexto del problema:** Ayuda a entender las características de los datos, su distribución, la relación entre variables y la estructura de la información, lo que mejora la toma de decisiones.
- **Mejorar la calidad del conjunto de datos:** Un EDA bien realizado puede llevar a una limpieza más efectiva del conjunto de datos, mejorando la precisión del modelo final.

Partes de un modelo de Machine Learning



Partes de un modelo de Machine Learning

In [4]:

```
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

Out[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Partes de un modelo de Machine Learning

Entradas del Modelo
Son las variables o datos de entrada que se utilizan para hacer predicciones

También conocidas como:

- Input
- Características (Features)
- Atributos
- Predictores
- Entradas
- Variables independientes
- Dimensiones
- X
- Probablemente más…

In [4]:
```
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

Out[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Partes de un modelo de Machine Learning

Salidas del Modelo
Son los valores o resultados que el modelo intenta predecir a partir de los datos de entrada

También conocidas como:

- Output
- Objetivo
- Respuesta
- Target
- Salida
- Variable dependiente
- Etiquetas
- Y
- Probablemente más…

In [4]:
```
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

Out[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Partes de un modelo de Machine Learning

Fila de datos (Input + Output)

Cada fila representa una observación o un caso específico dentro del conjunto de datos

También conocida como:

- Observación
- Punto de datos
- Registro
- Fila
- Probablemente más…

In [4]:

```
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

Out[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Partes de un modelo de Machine Learning

Etiquetas (en el contexto del aprendizaje supervisado)
Son los valores de las variables objetivo que el modelo intenta predecir

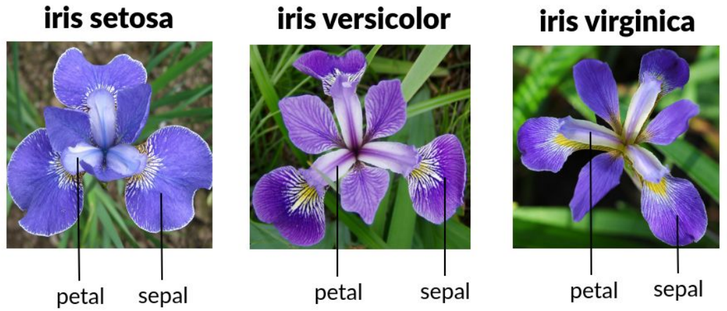
En este caso específico
las etiquetas son:

- Setosa
- Versicolor
- Virginica

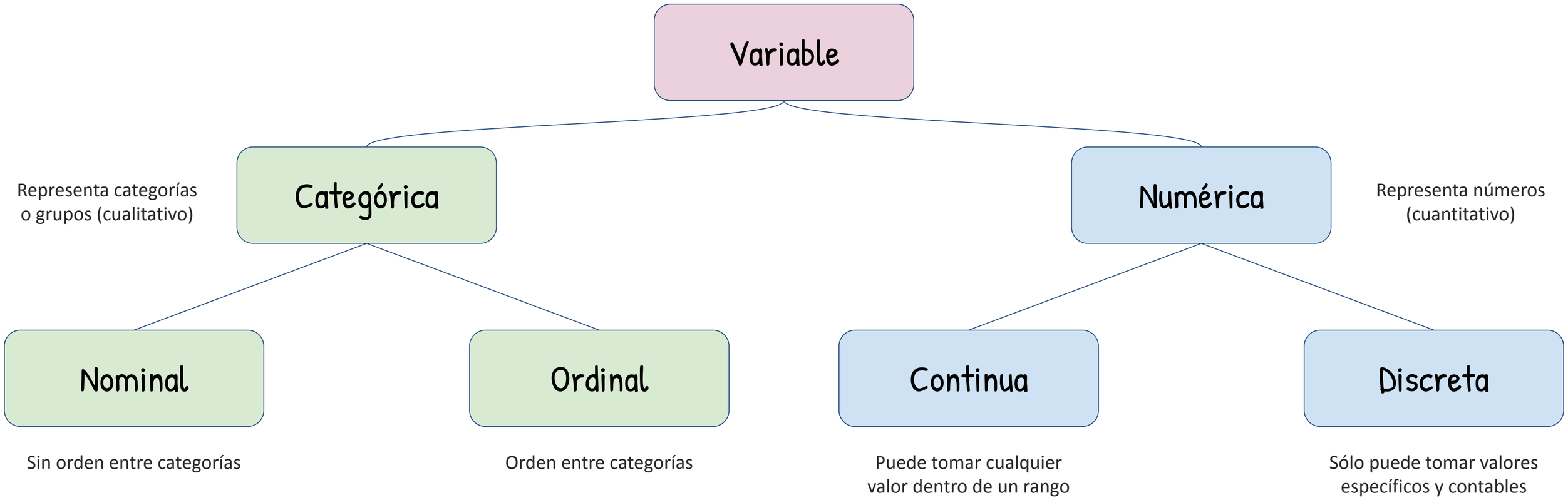
In [4]:
```
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

Out[4]:

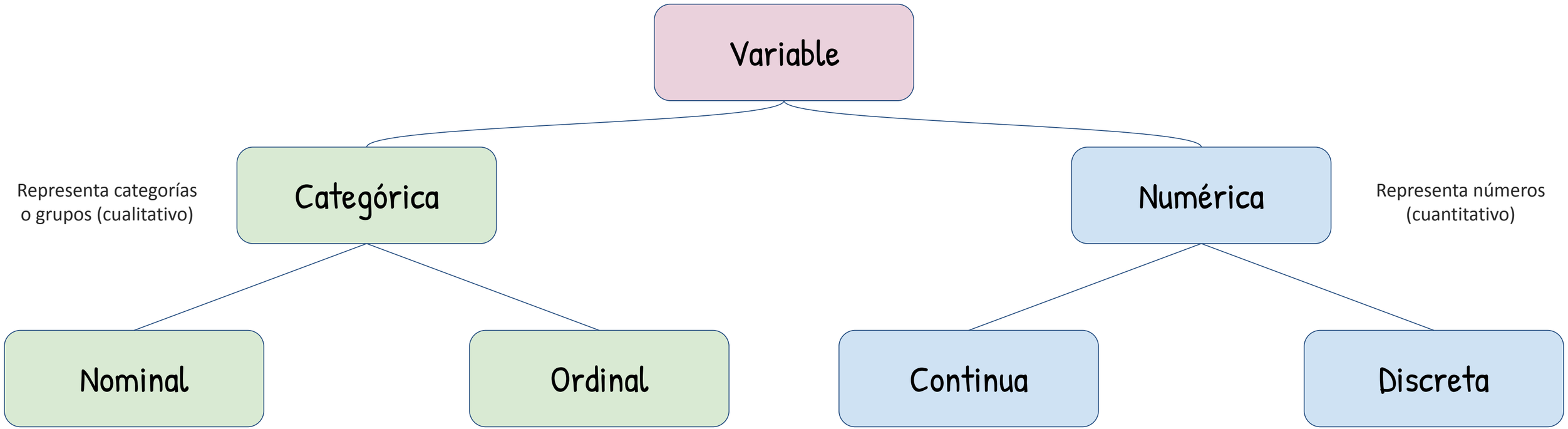
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa



Una variable puede ser categórica ó numérica



Una variable puede ser categórica ó numérica



Ejemplo • Color de ojos (azul/verde/cafés)	**Ejemplo** • Nivel de estudios (pregrado/máster/doctorado)	**Ejemplo** • Peso en kilos	**Ejemplo** • Número de vacas que un granjero tiene
Como variable objetivo, en ML es un modelo de **clasificación**	Como variable objetivo, en ML es un modelo de **Regresión** o **Clasificación**	Como variable objetivo, en ML es un modelo de **regresión**	Como variable objetivo, en ML es un modelo de **Regresión** o **Clasificación**

Estadísticas descriptivas

Métodos para resumir y describir las principales características de un conjunto de datos

Medidas de tendencia central

- **Media**
  Media aritmética de los datos
  Se calcula sumando todos los valores y dividiéndolos por el número de valores
- **Mediana**
  Valor medio cuando los puntos de datos están ordenados
- **Moda**
  El valor que aparece con más frecuencia en el conjunto de datos

Medidas de dispersión

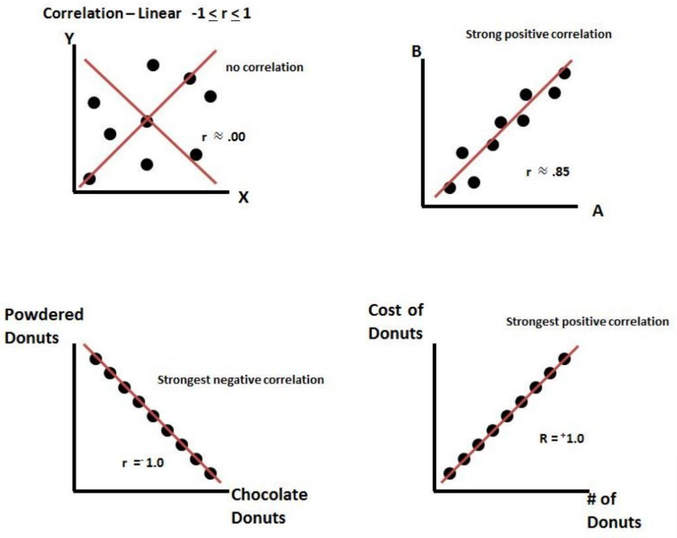
- **Rango**
  Diferencia entre los valores máximo y mínimo
- **Desviación estándar**
  Raíz cuadrada de la varianza, que da una idea de cuánto se desvían normalmente los puntos de datos de la media en las mismas unidades que los datos

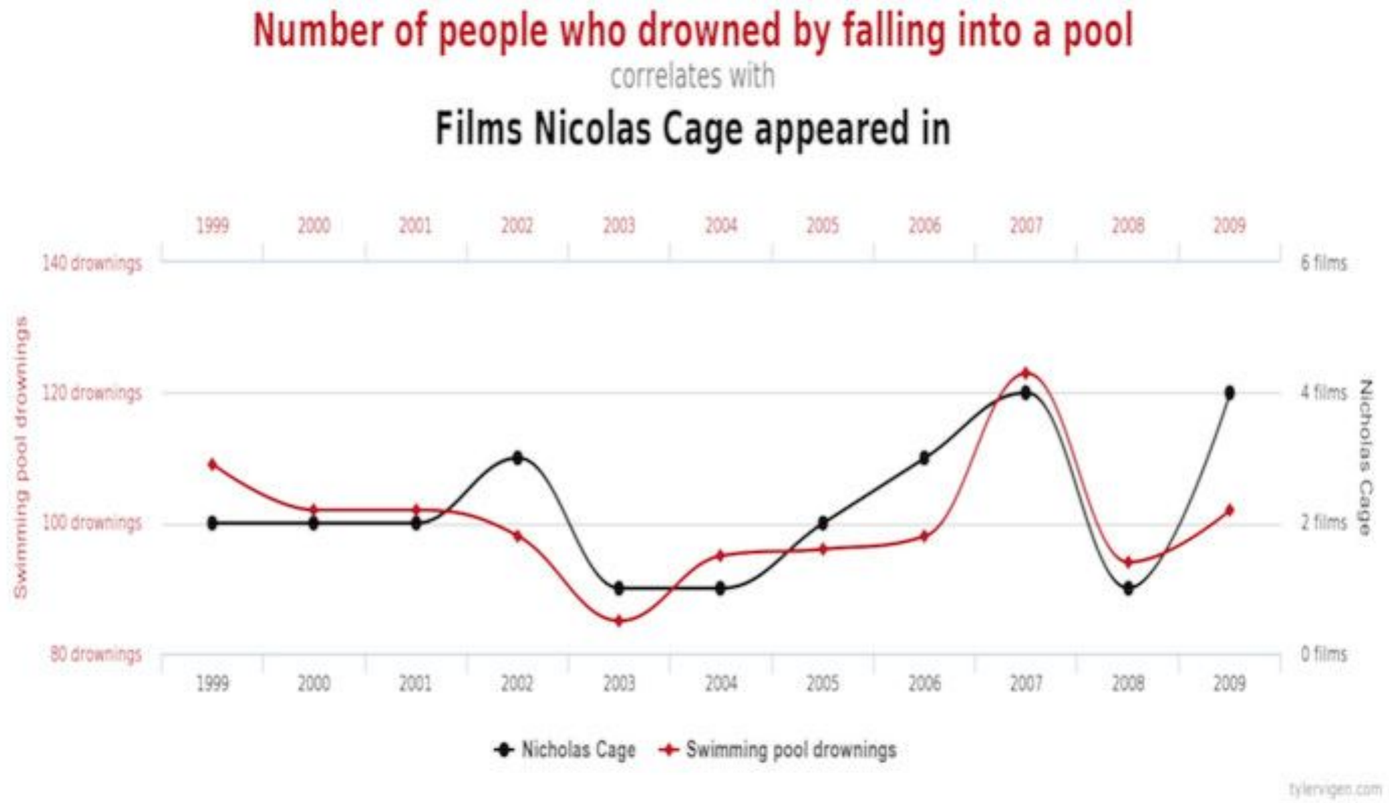
Correlación

La correlación es una medida estadística que cuantifica la **fuerza** y la **dirección** de la relación lineal entre dos variables. Oscila entre -1 y 1, donde:

- Un valor de 1 indica una correlación **positiva** perfecta, lo que significa que a medida que aumenta una variable, la otra aumenta proporcionalmente.
- Un valor de -1 indica una correlación **negativa** perfecta, lo que implica que a medida que aumenta una variable, la otra disminuye proporcionalmente.
- Un valor de 0 indica que **no existe correlación** lineal entre las variables.

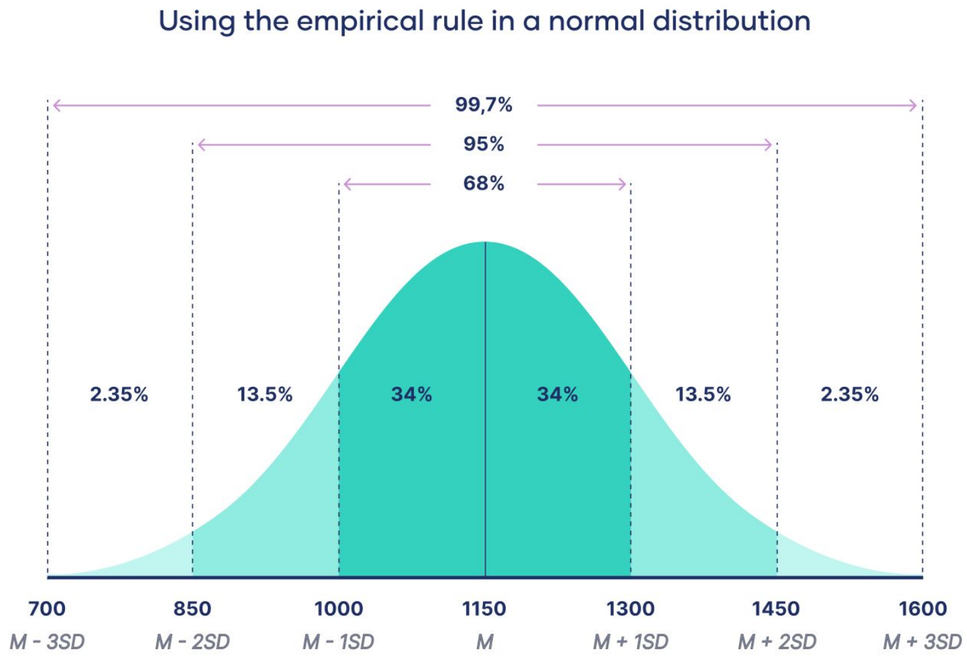
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$





¡La correlación NO implica causalidad!

Distribución normal (gaussiana)



Ejemplos:

- Alturas de un gran grupo de personas
- Medición de errores en experimentos científicos

- La media, la mediana y la moda son iguales.
- Aproximadamente el 68% de los valores se sitúan dentro de 1 desviación típica de la media, alrededor del 95% dentro de 2 desviaciones típicas y alrededor del 99,7% dentro de 3 desviaciones típicas. Esto se conoce como regla empírica.
- Está determinada por dos parámetros: la media (μ), que determina el centro de la distribución, y la desviación típica (σ), que determina la dispersión.

Notebook de hoy

https://colab.research.google.com/drive/1PJw_z5QhoimnpZk5Pqd5VODcoSjemW01?usp=sharing

Taller # 7

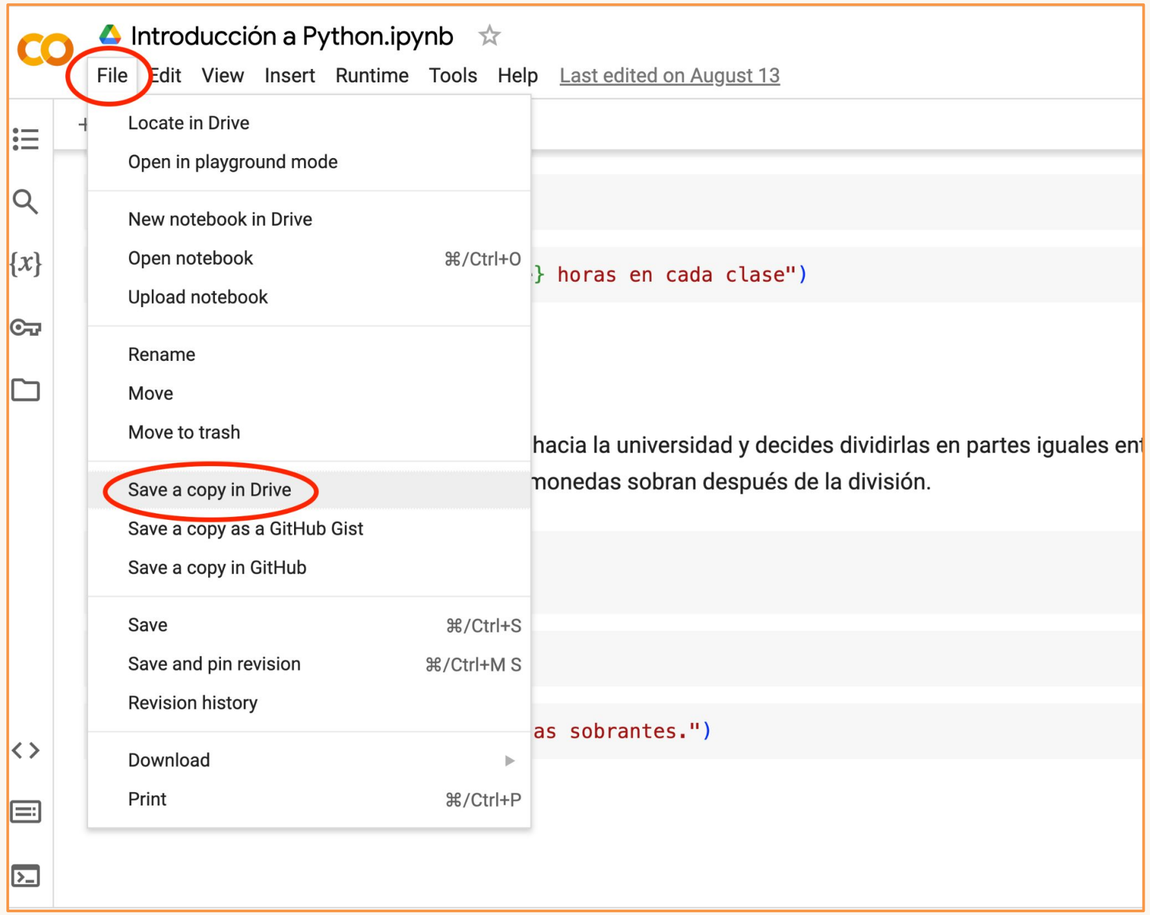
Exploración de datos

Encontrar un conjunto de datos y hacerle EDA en Python

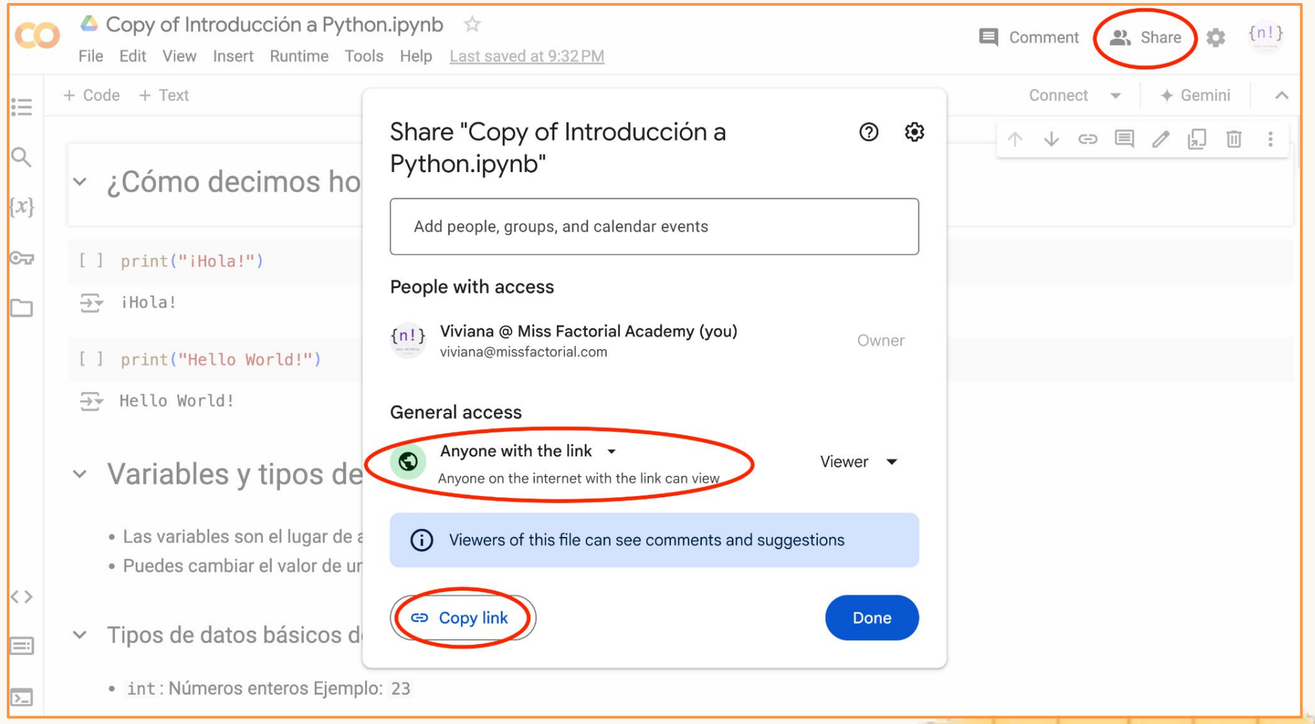
Fecha de entrega: Septiembre 23, 2024

Para enviar los talleres de código

- Hacer click en **archivo** → **guardar copia en mi Drive** para que les quede una copia en su cuenta, de lo contrario, los resultados no serán guardados.
- En la copia creada, hacer click en **compartir** , asegurarse que el enlace sea visible a **cualquier persona** , copiar el enlace y enviarlo.



vroberta@unicomfacauca.edu.co



¡Gracias!

¿Dudas? Email de la profe:
vroberta@unicomfacauca.edu.co
Página web del curso con toda la info:
https://github.com/vivianamarquez/unicomfacauca-ai-2024