Clase 01

Los temas para esta clase son :

1. Pandas básico

🤓 Conceptos clave para entendimiento de los datos

DataSet :

Se entiende un **dataset** como un conjunto de datos que contiene información representada de una manera especial y con un formato determinado.

ID	Nombre	País de origen	Fecha de nacimiento	Estatura (cm)	Código
0001	Andrés	Argentina	22/07/1961	173	A001
0002	Beatriz	Bolivia	05/01/1983	181	A002
0003	Carolina	Colombia	28/07/1991	177	B005
0004	Daniel	Perú	10/05/2005	189	E015

Data ó DataPoint :

Es el dato básico, es decir, un dato cualquiera que corresponda a una fila y una columna

ID	Nombre	País de origen	Fecha de nacimiento	Estatura (cm)	Código
0001	Andrés	Argentina	22/07/1961	173	A001
0002	Beatriz	Bolivia	05/01/1983	181	A002
0003	Carolina	Colombia	28/07/1991	177	B005
0004	Daniel	Perú	10/05/2005	189	E015

Pandas :

Es una librería creada sobre **Numpy** la cuál sirve y tiene gran utilidad para el entendimiento y análisis de los datos. Pandas al igual que Numpy implementa una nueva **estructura de datos** llamada ***Data Frame*.**

Se entiende un **DataFrame** como un arreglo de dos dimensiones que sirve para guardar información de un **dataset,** se puede entender fácilmente como una tabla de Excel en donde se tienen columnas que representas los **atributos de las variables** y filas que representan las **instancias de las variables.**

pd.DataFrame

	Código	Población	PIB	Área (km2)
Colombia	COL	50'372,424	343'177	1,141,748
Brasil	BRA	210'147,125	1'893'010	8,515,767
Argentina	ARG	44'938,712	443'249	2,780,400
Chile	CHL	19'107,216	308'588	756,096
Uruguay	URY	3'529,014	62'917	176,215
Ecuador	ECU	17'300,000	109'444	283,560
Venezuela	VEN	28'067,000	62'921	916,445
Bolivia	BOL	11'383,094	45'253	1,098,581
Perú	PER	32'912,915	290'217	1,285,216
Paraguay	PRY	7'152,703	42'826	406,752

Así mismo, dentro de los **DataFrame** existen otro **objeto** conocido con el nombre de **Series** que se refiere a las **columnas** o los "espacios" en donde podemos guardar los diferentes datos con diversos tipos de variables, es importante destacar que cada serie debe tener un **indice** que la diferencie de las otras.

pd.Series

Colombia	50'372,424
Brasil	210'147,125
Argentina	44'938,712
Chile	19'107,216
Uruguay	3'529,014
Ecuador	17'300,000
Venezuela	28'067,000
Bolivia	11'383,094
Perú	32'912,915
Paraguay	7'152,703

Argumentos útiles en Pandas

.name : En Pandas este atributo corresponde al nombre que tiene cada **Series** en el **DataFrame**, por defecto este nombre viene dado como **None** por lo cuál cada vez que se instancia una nueva **Series** se recomienda asignar un nombre para identificar.

.index : En Pandas se tiene este atributo para representar el **indice** que corresponde a una **Fila** ó **Columna** especifica.

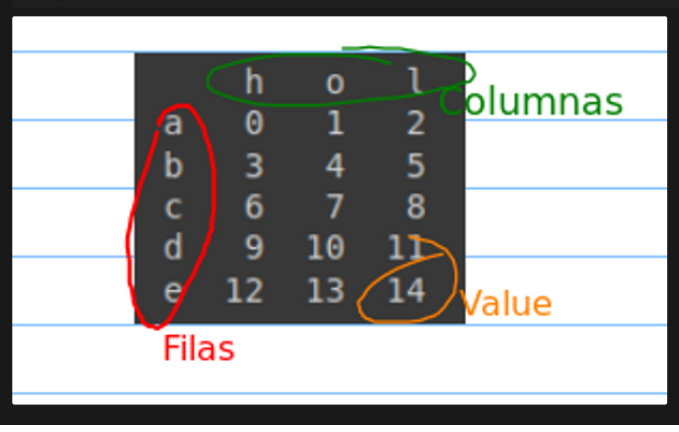
.columns : En Pandas este atributo representa las columnas que tiene un **DataFrame.**

.values : En Pandas este atributo arroja o menciona los valores que se tiene en un **DataFrame.**

.size : En Pandas este atributo nos arroja una constante de cuantos elementos existen por **Fila** ó **Columna.**

.shape : En Pandas este atributo nos arroja la dimensión que posee un **DataFrame.**

```
1 import pandas as pd
2
3 df = pd.DataFrame(np.reshape( np.arange( 0,15),(5,3) ), index = list('abcde'))
4 df.columns = list('hol')
5
6 print(df)
```

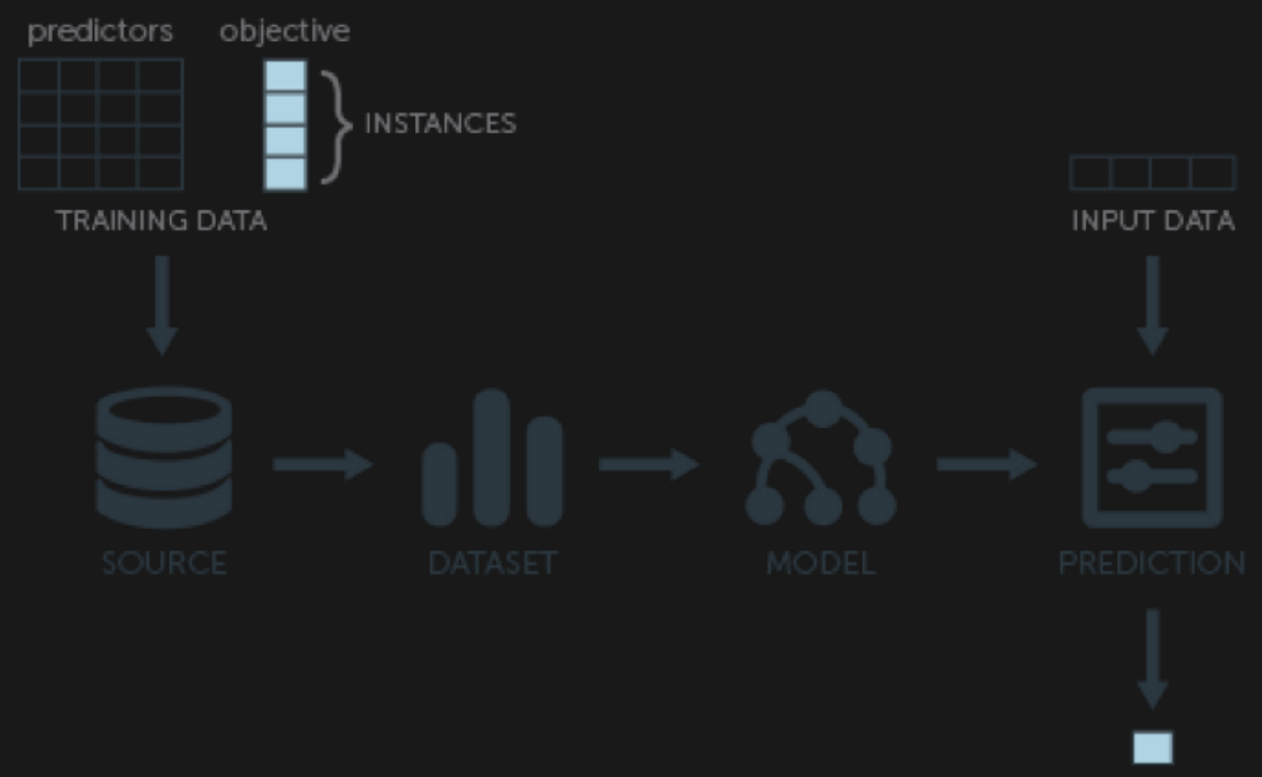


2. Introducción al ML (Machine Learning)

Machine Learning

Es aquella rama de la **Inteligencia Artificial** que se encarga de estudiar y aplicar las técnicas, algoritmos y métodos usados para hacer que un computador logre aprender de un conjunto de datos.

Para ello se crean modelos **descriptivos o predictivos (** Generar un conocimiento a partir de unos datos conocidos -Estadística Inferencial-, y *predictivo* genera un conocimiento con un conjunto de datos no conocido). Los modelos se pueden ver como la siguiente imagen :



Como existen modelos *predictivos* y *descriptivos,* existen modelos de **carácter supervisado y no supervisado.**

Aprendizaje Supervisado

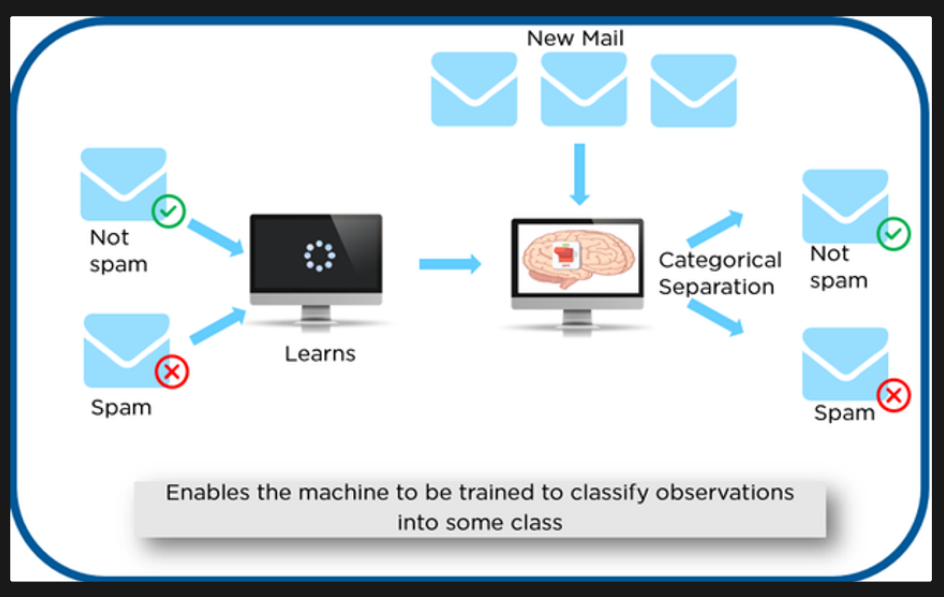
El aprendizaje supervisado se entiende como aquellos modelos que “aprenden” de manera supervisada, es decir, que las soluciones esperadas se encuentran previamente etiquetadas por un programa o por un ser humano.

Existen dos tipos de **aprendizaje supervisado:**

- **Clasificación:** Realiza una clasificación de acuerdo a unas etiquetas o características especificas.
- **Regresión:** Usa la estadística para realizar una predicción acerca de una variable

Ejemplos

- Reconocimiento de spam (El modelo sabe que características tiene un correo spam y realiza una ***predicción*** de los nuevos para realizar una ***clasificación***)

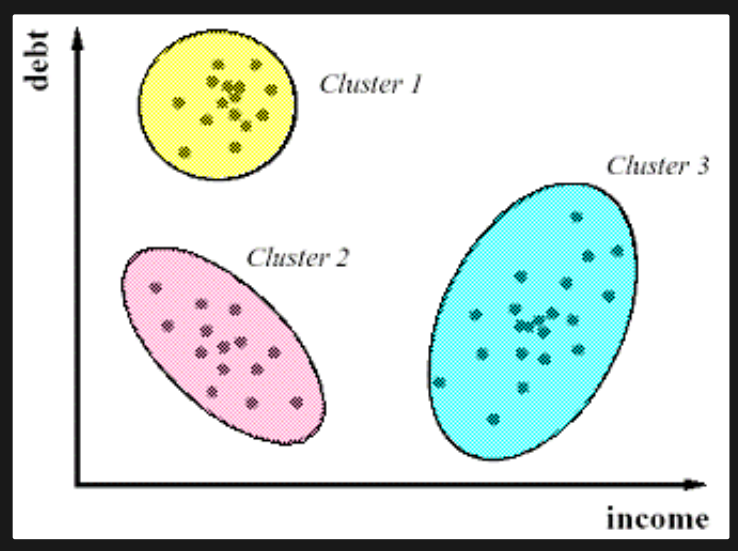


Aprendizaje No Supervisado

Este tipo de aprendizaje es muy diferente por que el modelo debe aprender por si solo ya que no existen ***etiquetas de valores esperados*** que el modelo pueda aprender. Es por ello que el modelo debe encontrar patrones o relaciones entre los datos para llegar a unas conclusiones.

Existen los tipo de aprendizaje de maquina :

- **Clustering:** Clasifica los datos en conjuntos de datos de acuerdo a una o varias características.



- **Asociación:** Asocia respuestas o salidas de acuerdo a los mismos datos suministrados. Por ejemplo, aquellas personas que compran un coche tienden a comprar un seguro para el mismo.

Problema de aprendizaje computacional	Descripción	Ejemplo	Tipo
Clasificación	Predecir la clase a la que pertenece un ejemplo.	Decidir si una imagen contiene un perro o un gato.	Supervisado
Regresión	Predecir un valor numérico en función de un ejemplo.	Predecir el precio de un inmueble.	Supervisado
Agrupamiento	Agrupar ejemplos similares.	Recomendación de películas.	No Supervisado
Reducción de dimensionalidad	Reducir la dimensión de los ejemplos de un conjunto de datos.	Convertir muchísimas características sin poder predictivo en pocas características con alto poder predictivo.	No Supervisado

Proceso de Machine Learning

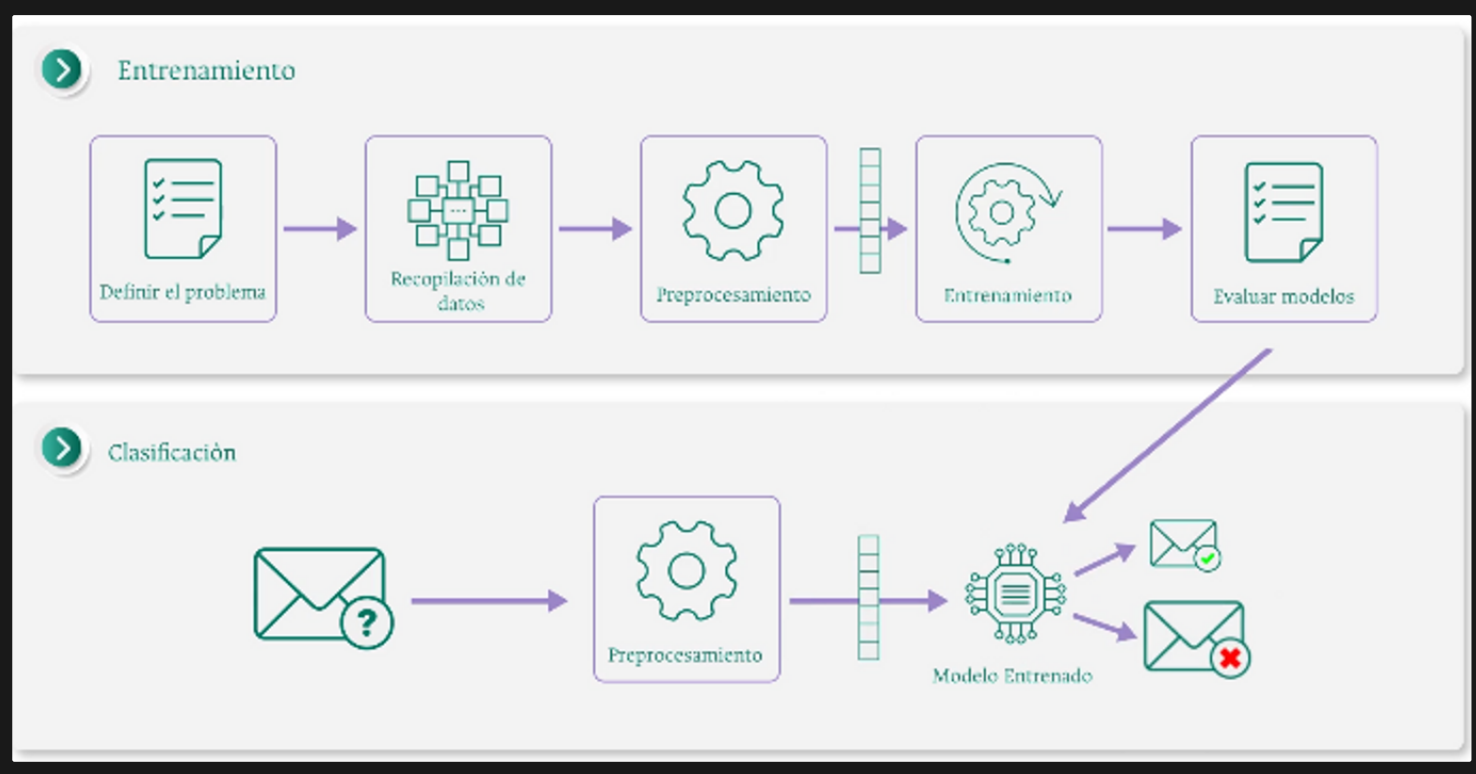
A continuación se muestran los pasos para realizar **Machine Learning**

- ▶ Definición de Problema
- ▶ Recolección de datos

- Preparación datos
- Entrenamiento de modelos
- Evaluación de modelos
- Predicciones de modelos

Ejemplo de Etapas de Machine Learning

A continuación se muestra un ejemplo para realizar la clasificación de correo en spam y no spam



1. Definición de problema:

   Se entiende el problema de clasificación de correos electronicos.

2. Recopilación de datos:

   Se recopilan los datos de correos eléctronicos que se encuentran clasificados en **spam** y **no spam**

3. Procesamiento de los datos:

   Se transforman los datos de String a valor numérico asignandolo a una Matriz dada

4. Entrenamiento:

   Se entrena el modelo de aprendizaje supervisado y se vuelve a revisas constantemente para posteriormente evaluarlo

5. Producción:

   El modelo creado recibe nuevos datos y se encarga de realizar una predicción para clasificarlos como **spam** o **no spam**

Preprocesamiento de Variables

Entendemos el preprocesamiento de los datos como todos aquellos métodos y procesos realizados para **limpiar** los datos y garantizar una calidad **optima** de los mismos para posteriormente realizar un entrenamiento o análisis adecuado.

Existen varias maneras de realizar un preprocesamiento de los datos, la primera parte de ***data cleaning*** se mencionó en unidades anteriores. Sin embargo, existen a su vez diversas formas que son importantes aplicar. Es el caso de:

Estandarización o Normalización

Esta técnica estadística nos permite darle un formato especifico a conjuntos de datos que tienen medidas diferentes (cm y m) para poder realizar una comparación entre los mismos. Así mismo, son usadas para convertir una distribución de conjuntos en una distribución conocida como la **distribución normal.**

Matemáticamente la estandarización más conocida es :

$$X' = \frac{X - \mu}{\sigma}$$

Donde:

- μ : Media aritmética de los datos.
- σ : Desviación estándar de los datos.

Ejemplo:

Tenemos un **dataset** con una distribución similar a una ***distribución normal*** **.** Y queremos hallar la probabilidad de encontrar un valor en un intervalo dado, como la distribución no es estándar lo que debemos realizar es estandilizar toda la distribución y a partir de esto aplicar una ***distribución normal estándar*** y encontrar la probabilidad que anteriormente nos preguntaban.