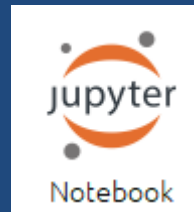


Indice

- ❖ Instalación de herramientas para el aprendizaje automático
 - Ejemplo de aplicación de aprendizaje supervisado / regresión
- ❖ Clasificación de sistemas de aprendizaje automático (II)
 - Aprendizaje no supervisado
 - Aprendizaje semisupervisado
 - Aprendizaje por refuerzo
 - Aprendizaje por lotes
 - Aprendizaje Online
 - Basado en instancias
 - Basado en modelos
- ❖ Las 7 Fases del Proceso de Machine Learning
- ❖ Preprocesamiento - Contextualización
- ❖ Preprocesamiento
 - Estadística descriptiva
 - Distribución normal
 - Outliers
 - Observaciones influyentes
 - Escalamiento (Estandarización)
 - Selección de variables y Ponderación de variables
- ❖ Cuaderno demo: Panda, Numpy, Matplotlib y carga de ficheros

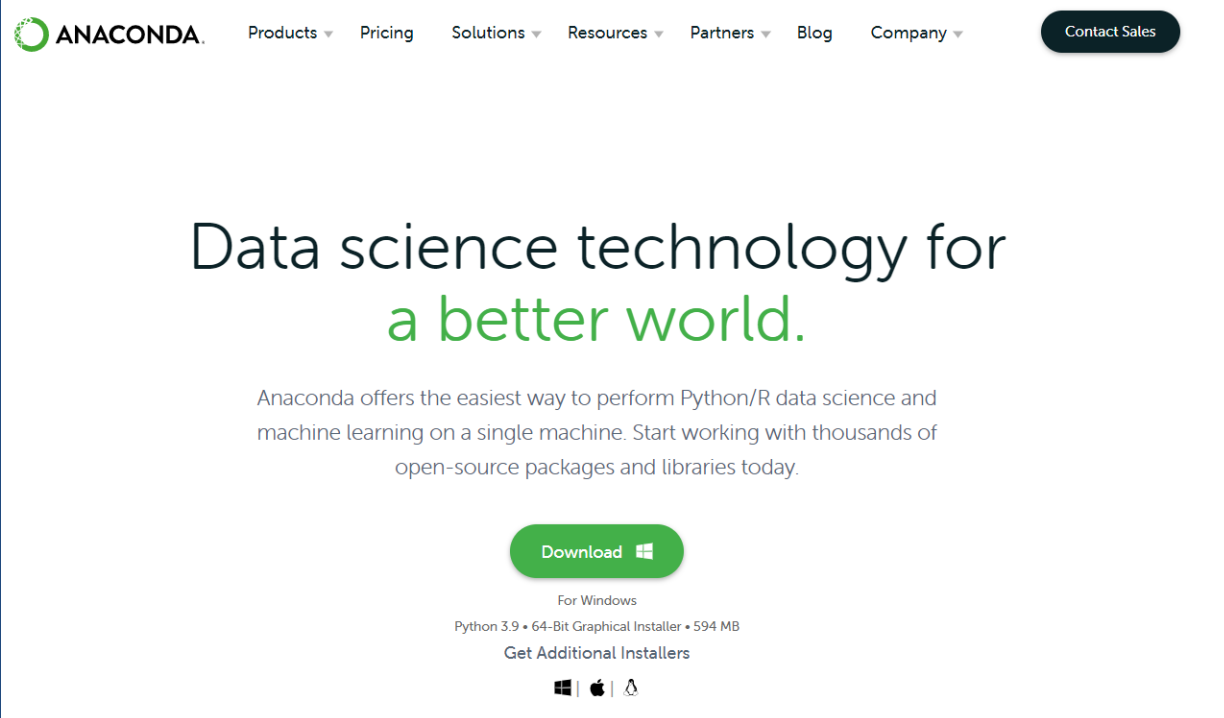
Herramientas





UT2 – Técnicas y herramientas de aprendizaje automático

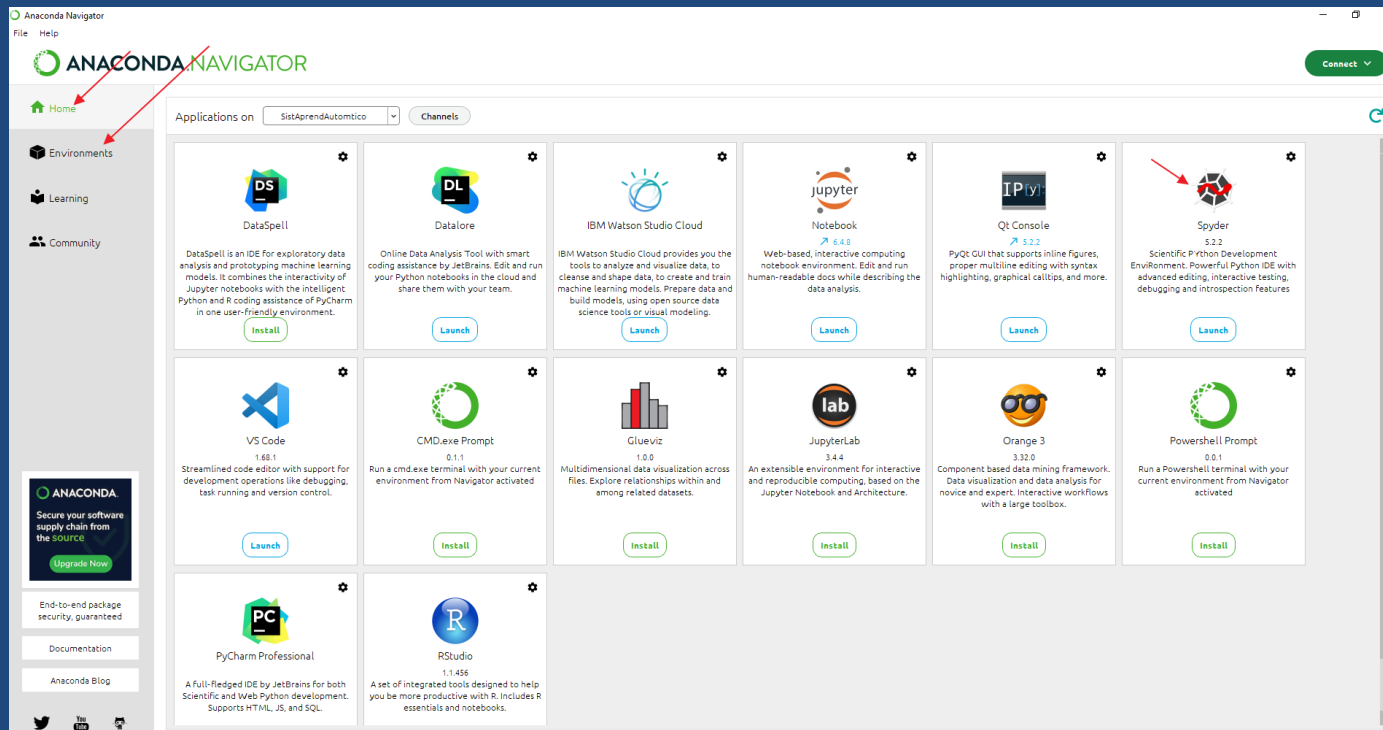
Instalación de herramientas para el aprendizaje automático



The screenshot shows the Anaconda website homepage. At the top, there is a navigation bar with the Anaconda logo, links for Products, Pricing, Solutions, Resources, Partners, Blog, and Company, and a Contact Sales button. The main content area features the headline "Data science technology for a better world." with "a better world." in green. Below this, a paragraph states: "Anaconda offers the easiest way to perform Python/R data science and machine learning on a single machine. Start working with thousands of open-source packages and libraries today." A prominent green "Download" button with a Windows icon is centered. Below the button, it specifies "For Windows" and "Python 3.9 • 64-Bit Graphical Installer • 594 MB". A link "Get Additional Installers" is provided, followed by icons for Windows, macOS, and Linux.

url: <https://www.anaconda.com/>

Instalación de herramientas para el aprendizaje automático



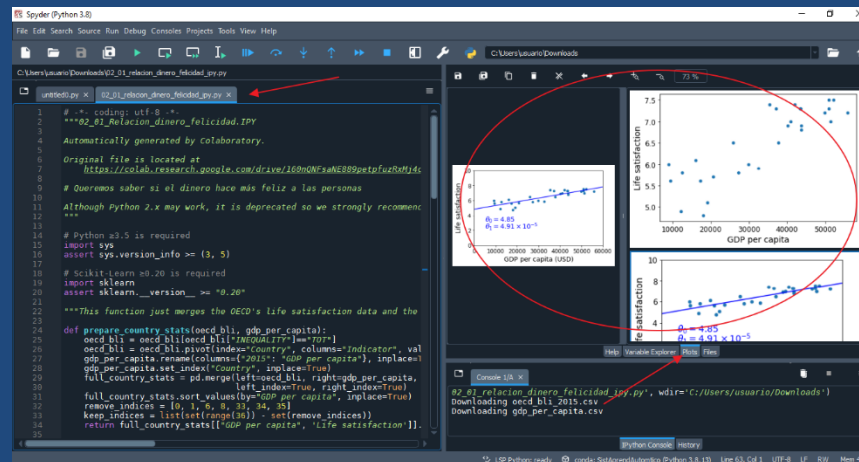
Comentar:

- Instalación de paquetes
- Posibilidad de actualizar paquetes desde la consola

Instalación de herramientas para el aprendizaje automático



Para hacer nuestra primera prueba descargamos el cuaderno (con formato py) 02_01_Relacion_dinero_felicidad.IPY y lo ejecutamos en **Spyder**. Cuando lo ejecutemos nos pedirá que instalemos unas librerías ...

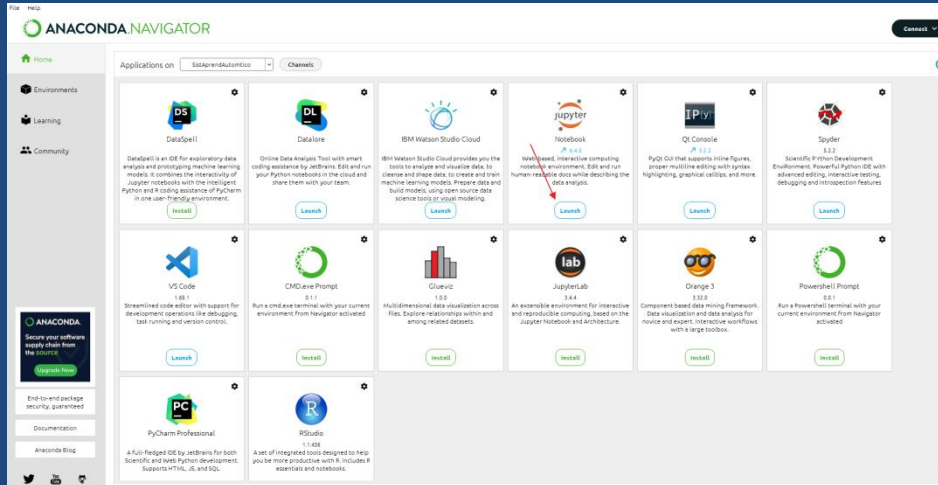


https://colab.research.google.com/drive/1oXHIXIAegsV3VSEMKfDeOlgbZLm3_uhL?usp=sharing

El objetivos de este ejemplo:

- Ver que se pueden descargar los cuadernos en formato .py y lo podemos ejecutar en Spyder. Creando incluso los directorios y fichero de trabajo en local.
- Aunque en este ejemplo utilizamos un modelo de regresión lineal, lo explicaremos con detalle más adelante (Diapo 11). De momento lo descargamos y ejecutamos en Spyder y debería ejecutarse sin dificultades.

Instalación de herramientas para el aprendizaje automático



Si Anaconda muestra una última versión de Jupyter entonces actualizar: En la tuerca superior derecha + Update Application

- Interfaz web de código abierto que permite la inclusión de texto, imágenes, video y audio, así como la ejecución de código a través del navegador en múltiples lenguajes.
- Incluye por defecto únicamente el núcleo de cálculo Python
- Su nombre surge al unir 3 de los lenguajes de programación de código abierto más utilizados en el ámbito científico: **Ju**-lia, **Py**-thon y **R**)
- Extensión de los fichero Jupyter Notebook: ipynb

Título: .ipynb Extensión de archivo

Url: <https://abrirarchivos.info/extension/ipynb>

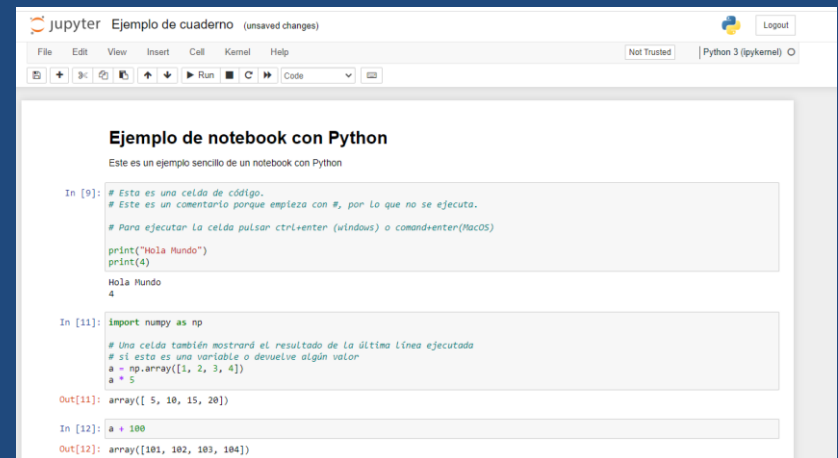
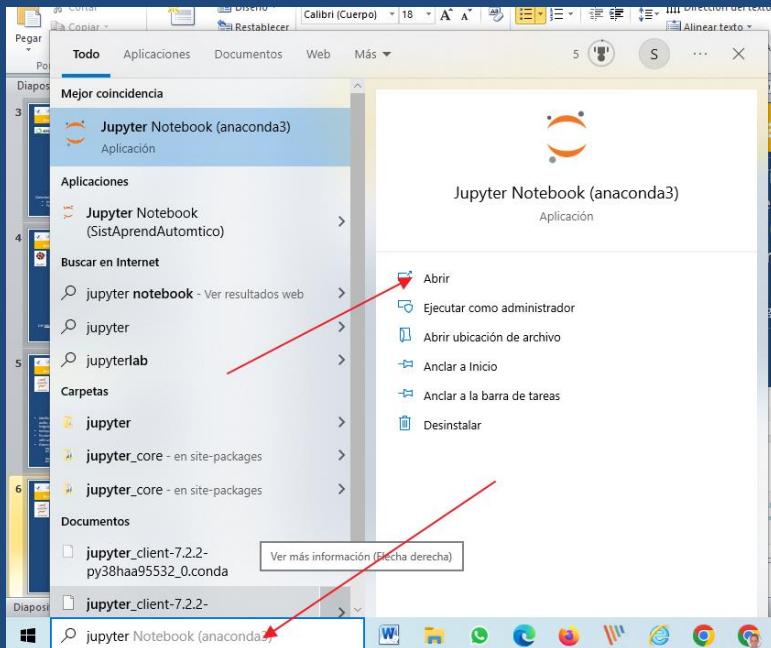
Título: Jupyter: Data Science aplicada

Url: <https://www.paradigmadigital.com/dev/jupyter-data-science-aplicada/>

Instalación de herramientas para el aprendizaje automático



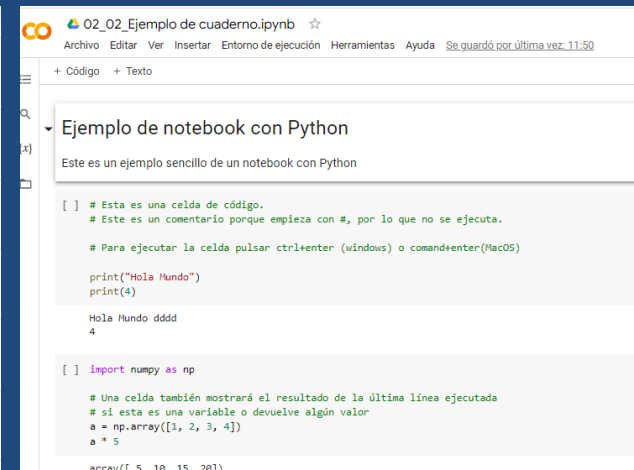
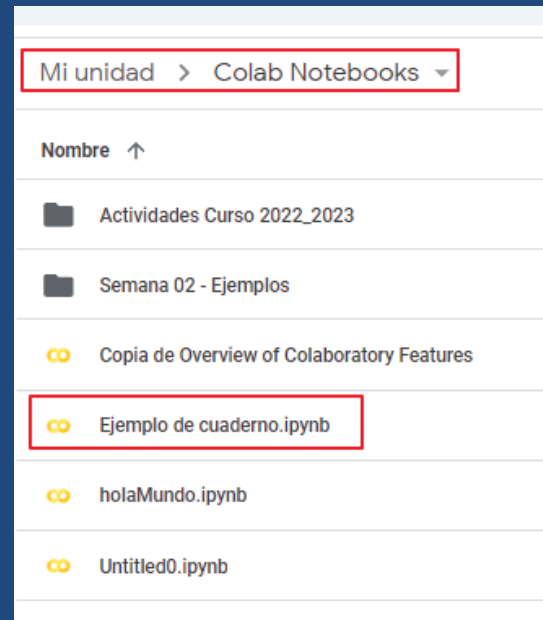
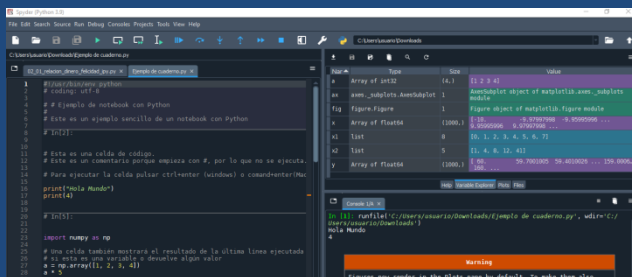
- Lanzar Jupyter Notebook en **Chrome de Windows**: En la barra de búsqueda de Windows 11 escribir **Jupyter** y entonces se ejecutará sobre el Chrome.
- Crear un directorio de pruebas desde Jupyter Notebook. En Windows las carpetas se crean en el directorio **C:\Usuarios\usuario**
- En dicho directorio descargar el archivo de prueba de la plataforma: **02_02_Ejemplo de cuaderno.ipynb**
- Url: <https://colab.research.google.com/drive/1BDkADPuglw70GypkMU90zg4l4FAMC409?usp=sharing>



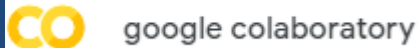
Instalación de herramientas para el aprendizaje automático



- Analizar los contenidos del Jupyter Notebook y realizar modificaciones.
- Mostrar cómo exportar a formato python desde Jupyter notebook y abrirlo en Spyder.
- Subir el cuaderno a Google Colab. (Ver previamente instalación de Google Colab)
- Tener en cuenta que al subirlo se guarda directamente en la carpeta Mi unidad/Colab Notebooks.
- Trasladarlo a la carpeta que se desee después de subirlo, realizar pruebas y ver cómo se pueden monitorizar las variables.



Instalación de herramientas para el aprendizaje automático

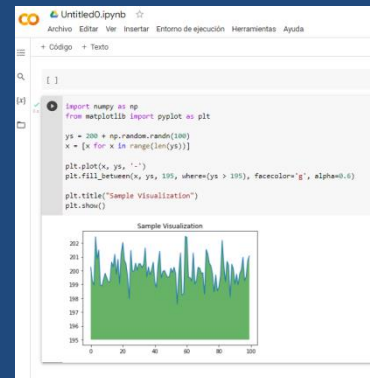
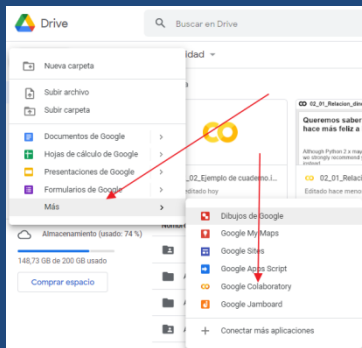


1.- Instalación de Google Colab y creación de un cuaderno.

2.- Copiar celda de prueba de la web de Google colab o bien abrir el ejemplo

02_03_Ejemplo_plt.ipynb

https://colab.research.google.com/drive/1JOukF_DP4AZ6l2fDGzaeaOnZ-8XxZEoJ?usp=sharing



Título: Te damos la bienvenida a Colaboratory

Url: <https://colab.research.google.com/?hl=es>

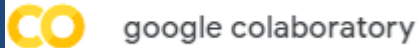
Título: Google Colab – Preguntas frecuentes -- Para conocer los límites

Url: <https://research.google.com/colaboratory/faq.html#gpu-availability>

Título: Rendimiento de CPU vs GPU vs CPU y diferencias discutidas

Url: <https://br.atsit.in/es/?p=287622>

Instalación de herramientas para el aprendizaje automático



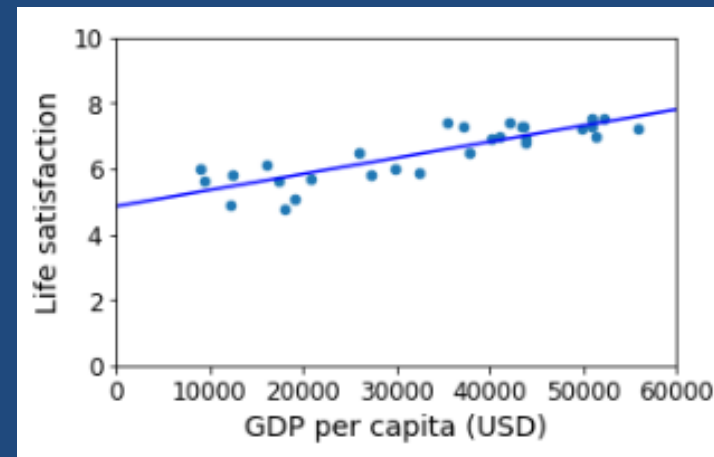
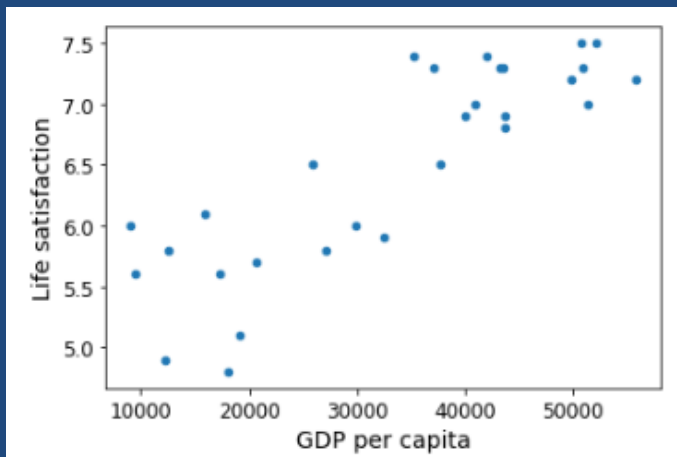
Ejemplo de aplicación de aprendizaje supervisado / regresión

1.- Acceder al siguiente cuaderno para ver un ejemplo de regresión:

Nombre: 02_01_Relacion_dinero_felicidad.IPY

https://colab.research.google.com/drive/1oXHIXIAegsV3VSEMKfDeOlgbZLm3_uhL?usp=sharing

2.- Realizamos nuestras primeras pruebas y conoceremos en código el concepto de parámetros e hiperparámetros..



Clasificación de sistemas de aprendizaje automático (II)

Aprendizaje supervisado y no supervisado

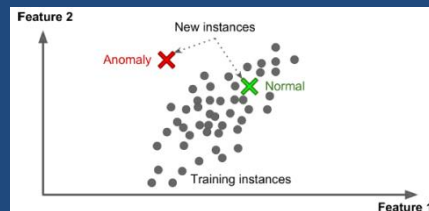
Tipo de aprendizaje

**Aprendizaje no
supervisado**

Tipo de problema

**Detección de anomalías/novedades
Reglas de asociación**

- Ejemplos:
 - Detección de operaciones con tarjetas de créditos para prevenir el fraude.
 - Detectar valores anómalos/atípicos antes de introducirlos en un algoritmo de aprendizaje.
 - Detectar novedades en los datos que sean muy diferentes a los datos de entrenamiento.
 - Explorar cantidades enormes de datos y encontrar asociaciones. Por ejemplo, analizar compras de productos y definir perfiles de clientes.



Clasificación de sistemas de aprendizaje automático (II)

Aprendizaje supervisado y no supervisado

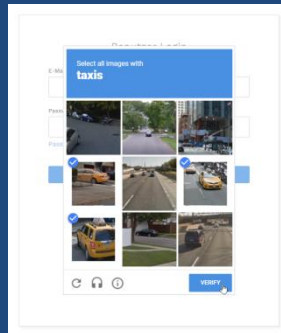
Tipo de aprendizaje

**Aprendizaje
semisupervisado**

Tipo de problema

Etiquetado de datos e imágenes

- El etiquetado de datos lleva mucho tiempo y es muy costoso
- Ejemplos:
 - Google Captcha
 - Google Fotos



Título: Las etiquetas de tus fotografías de Google Fotos servirán para entrenar la inteligencia artificial del servicio

Url: <https://www.xatakandroid.com/aplicaciones-android/etiquetas-tus-fotografias-google-fotos-serviran-para-entrenar-inteligencia-artificial-servicio>

Clasificación de sistemas de aprendizaje automático (II)

Aprendizaje supervisado y no supervisado

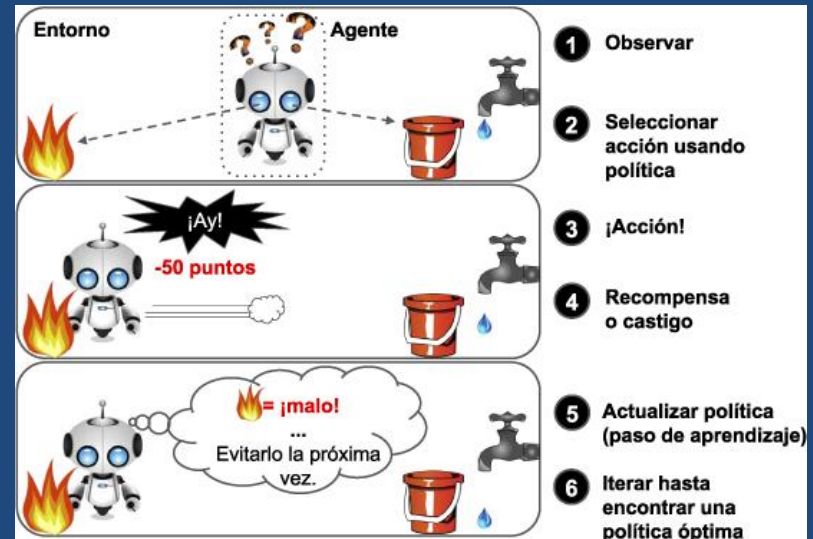
Tipo de aprendizaje

**Aprendizaje por
refuerzo**

Tipo de problema

**Aprender por sí mismo cuál es la
mejor estrategia**

- El sistema observa el entorno, seleccionar y realizar acciones para recibir **recompensas** o **castigos**. Debe aprender por sí mismo cuál es la mejor estrategia y así obtener la mayor recompensa a lo largo del tiempo.
- Ejemplo:
 - Los robots aprendiendo a andar.
 - Juego GO



Clasificación de sistemas de aprendizaje automático (II)

Aprendizaje por Lotes y aprendizaje Online:

Si el sistema puede o no aprender de forma gradual a partir de un flujo de datos.

Tipo de aprendizaje

Aprendizaje por lotes

Tipo de problema

Aprendizaje utilizando un conjunto de datos muy grande que se actualiza periódicamente

- Es un sistema que aprende de forma gradual: utilizando todos los datos disponibles.
- Requiere muchos recursos y tiempo, por lo que el aprendizaje se suele realizar **offline**
- Aprendizaje offline: primero se estrena el sistema y después se lanza a producción.
- Estos sistemas se pueden automatizar: De manera periódica se actualizan los datos de entrenamiento, se entrena una nueva versión y se sube a producción.
- Ejemplo:
 - Detección de SPAM – **Muy instructivo ejemplo del artículo.**
(procesadores del lenguaje natural)



Título: Inteligencia artificial contra el spam

Url: <https://www.computerworld.es/archive/inteligencia-artificial-contra-el-spam>

Clasificación de sistemas de aprendizaje automático (II)

Aprendizaje por Lotes y aprendizaje Online

Si el sistema puede o no aprender de forma gradual a partir de un flujo de datos.

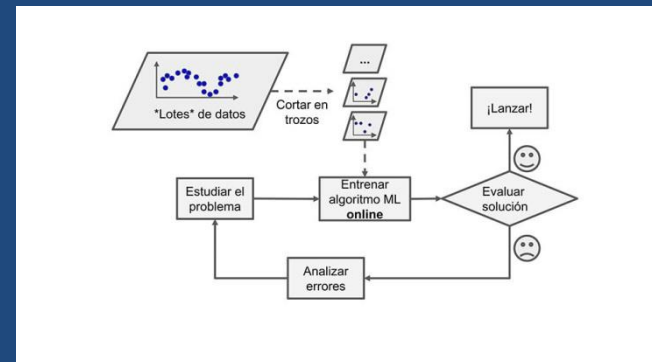
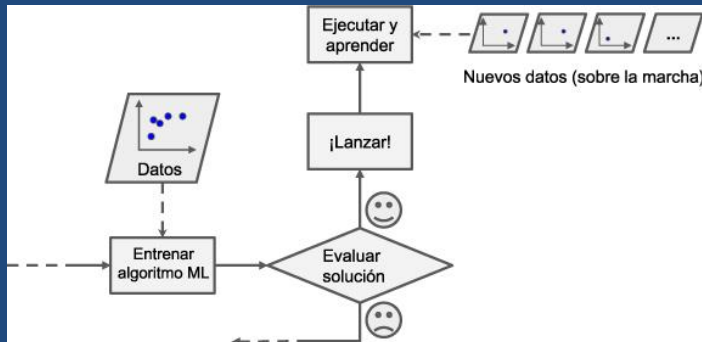
Tipo de aprendizaje

Aprendizaje Online

Tipo de problema

Necesitan adaptarse al cambio con rapidez

- Es una buena opción con recursos limitados
- Como inconveniente, es necesario supervisarlo continuamente por si recibe datos malos (por sensores en mal estado o datos erróneos provocados de manera intencionada) de entrada





Clasificación de sistemas de aprendizaje automático (II)

Aprendizaje por Instancias y Modelos

La mayoría de los sistemas de aprendizaje de IA funcionan por predicciones. Es decir se entrenan con una serie de datos y a partir de su aprendizaje, se intenta obtener buenas /predicciones/decisiones/conclusiones utilizando nuevos datos: es lo que se conoce como **generalización**.

Para la generalización hay dos enfoques. Aprendizaje basado en **Instancias** y aprendizaje basado en **Modelos**.

Clasificación (II) de sistemas de aprendizaje automático

Aprendizaje por Instancias y Modelos

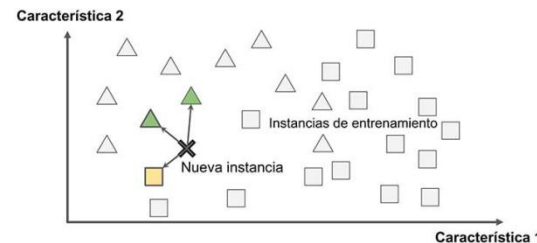
Tipo de aprendizaje

Basado en instancias

Tipo de problema

Obtener predicciones por similitudes (memoria)

- El sistema aprende los ejemplos de **memoria** después generaliza a nuevos casos aplicando una medida de solicitud (comparando con ejemplos previos de manera muy básica)
- Ejemplo: Para detectar correo SPAM se marcan los correos que sabemos que son spam y después los nuevos correos se comparan con los marcados previamente y un función de un grado de similitud se considerarán si son o no spam (por ejemplo: núm de palabras similares, etc.).



Clasificación de sistemas de aprendizaje automático (II)

Aprendizaje por Instancias y Modelos

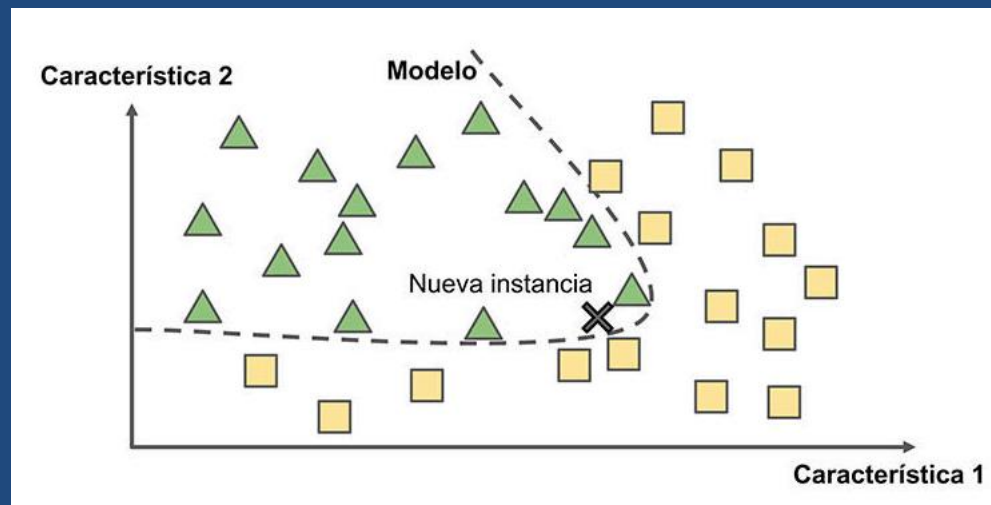
Tipo de aprendizaje

Basado en modelos

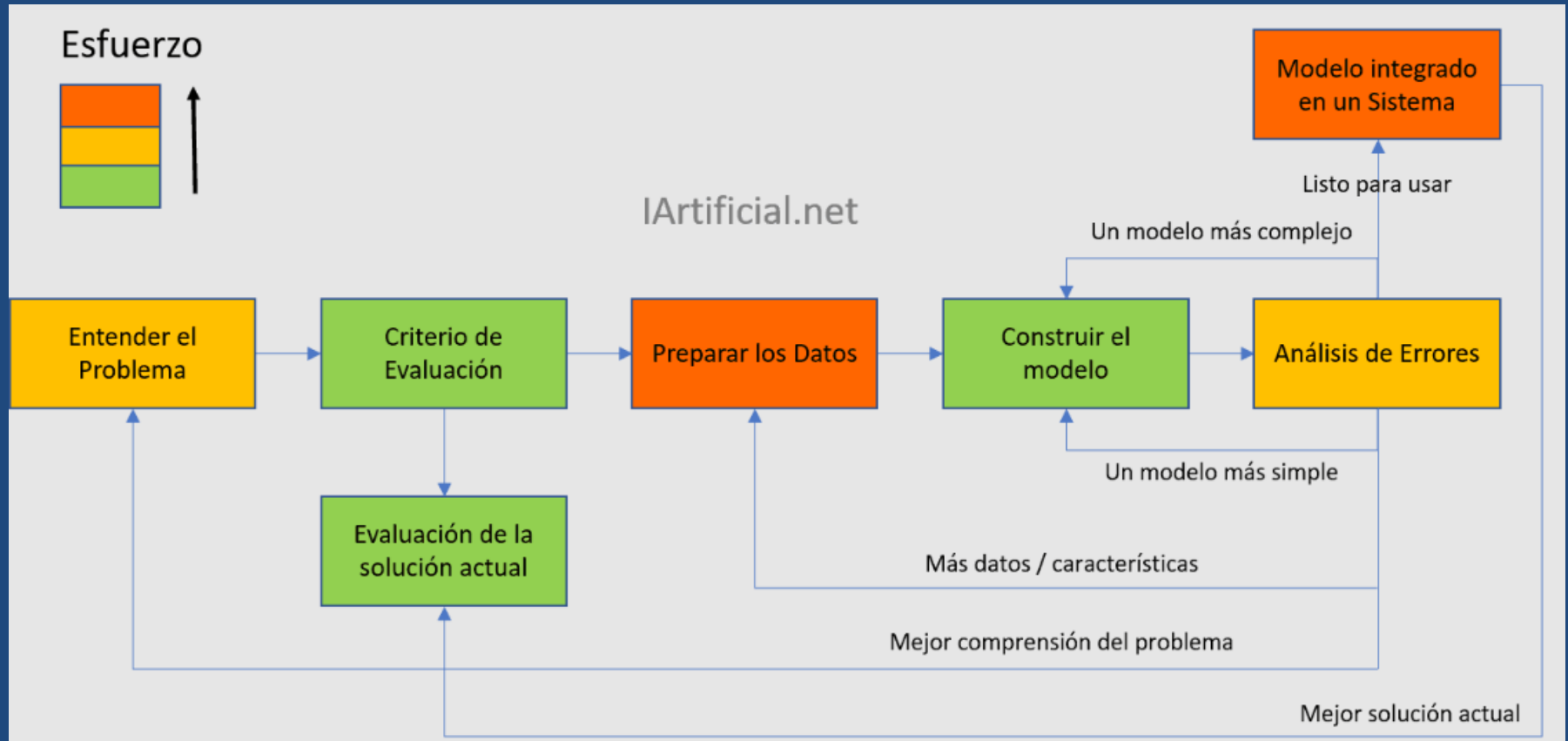
Tipo de problema

Obtener predicciones por similitudes asociadas a un modelo matemático

- Crear un modelo a partir de un conjunto de datos y después se utiliza para hacer predicciones.
- Ejemplo: algoritmo para saber si el dinero hace más feliz a las personas.



Las 7 Fases del Proceso de Machine Learning



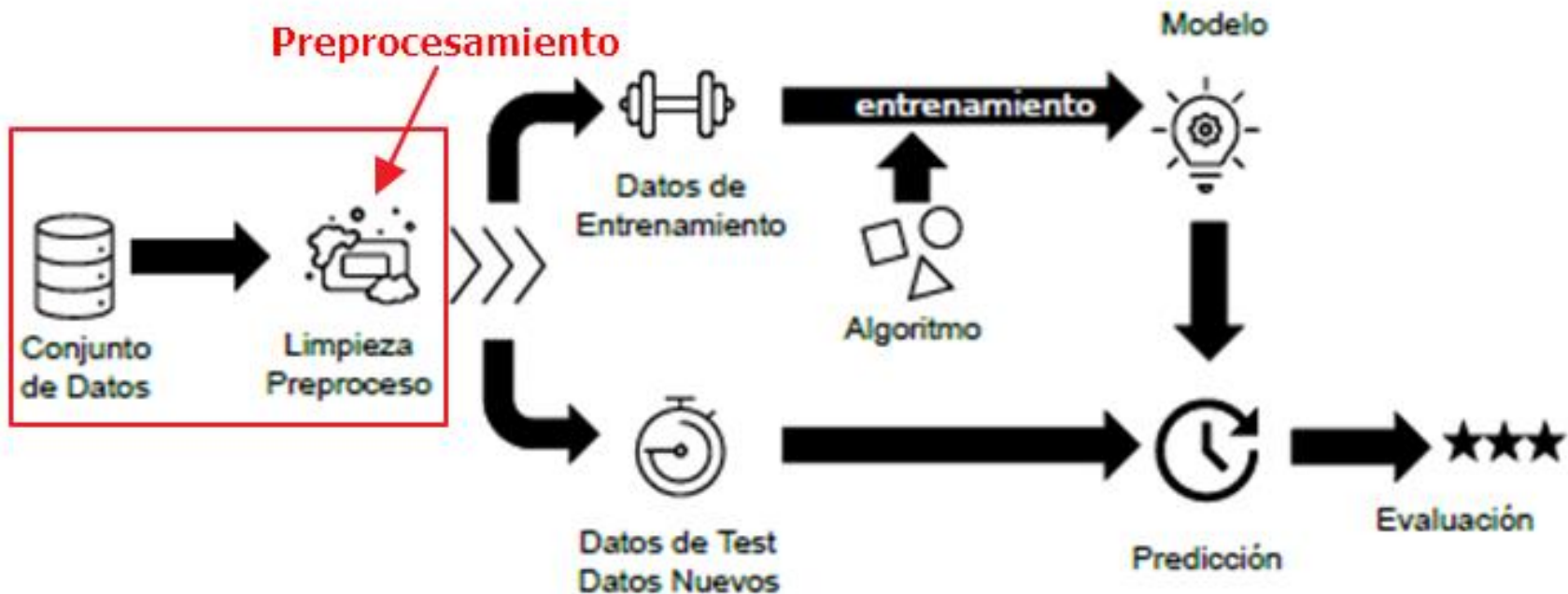
Título: Las 7 Fases del Proceso de Machine Learning

Url: <https://www.iartificial.net/fases-del-proceso-de-machine-learning/>

Preprocesamiento - Contextualización

A continuación se muestra un esquema básico de los pasos a realizar en el aprendizaje supervisado. Profundizaremos en la **UT3 - Algoritmos y herramientas para el aprendizaje supervisado**

Podemos observar el momento en el que se realiza el preprocesamiento:



Preprocesamiento

Definición: Conjunto de tareas encaminadas a la preparación de los datos.

Las tareas a realizar son las siguientes:

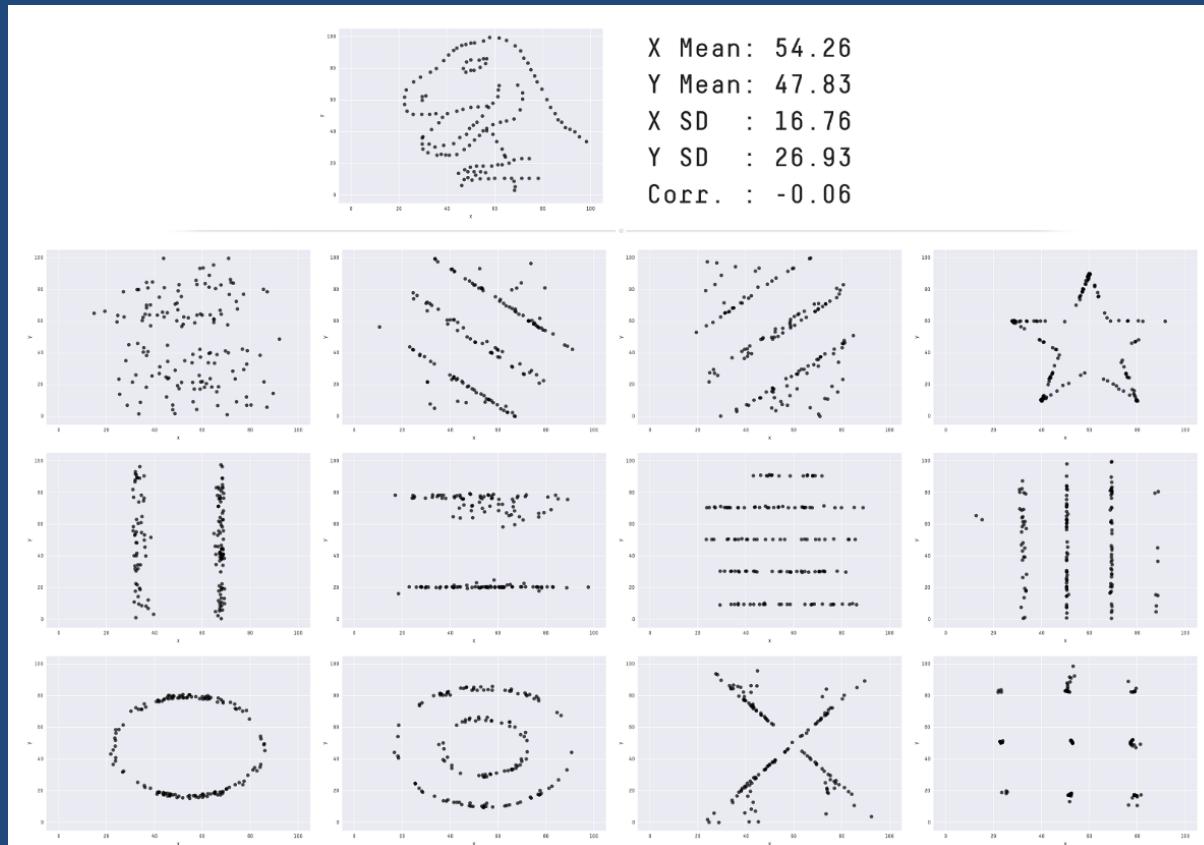
- Trabajar con una copia de los datos, poner a buen recaudo los datos originales.
- Exploración **visual** y **estadística** (outliers, observaciones influyentes, etc.) de los datos con el objetivo de detectar y analizar incoherencias, o bien, seleccionar los datos que realmente son útiles en la solución que se busca.
- Detectar valores anómalos que pueden influir fuertemente en las conclusiones del análisis → **outliers**.
- Determinar qué se hace con los registros incompletos → **missing values**:
 - Completarlos de alguna forma.
 - Eliminar las filas completas.
- Analizar/observar la magnitud con las que se miden las diferentes variables, pudiendo influir en el análisis de los datos por lo que en ocasiones conviene la **estandarización** de datos para convertirlos todos a una **misma escala**.
- Transformar datos ordinales en valores numéricos.
- Si el tamaño de los datos es excesivamente grande, se requiere la reducción de la dimensionalidad y/o selección de variables significativas. Ver variables del ejemplo del programa de Ingresos vs Felicidad.
- Intentar crear funciones para realizar la transformación de los datos.

Preprocesamiento

Por qué es tan importante visualizar los datos:

Título: Las mismas estadísticas, diferentes gráficos (Traducido por Google Translate)

Url: <https://www.research.autodesk.com/publications/same-stats-different-graphs/>

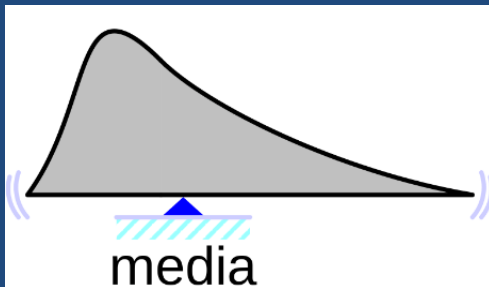


Preprocesamiento: Estadística descriptiva

Medidas de centralidad

Media

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

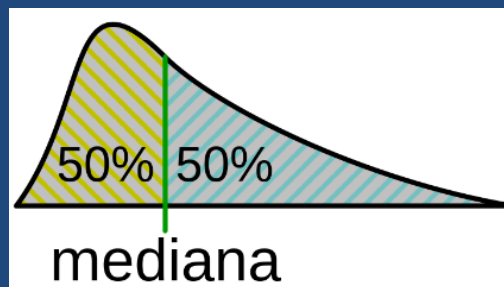


Mediana

$$n \text{ es impar} \Rightarrow Me = x_{(\frac{n+1}{2})}$$

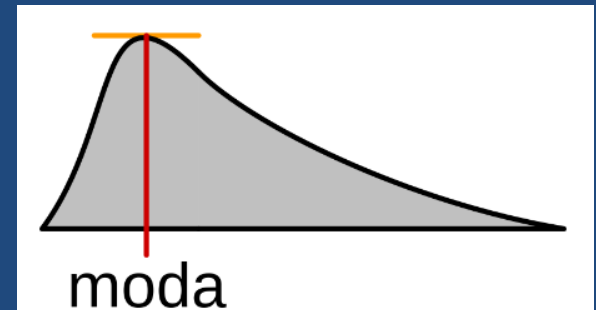
$$n \text{ es par} \Rightarrow Me = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

Es el valor central cuando los datos se ordenan



Moda

Para variables cualitativas es el valor que más se repite.
(1 Andorra, 2 Angola, etc...)



Título: Mediana (estadística)

Url: [https://es.wikipedia.org/wiki/Mediana %28estad%C3%ADstica%29](https://es.wikipedia.org/wiki/Mediana_%28estad%C3%ADstica%29)

Preprocesamiento: Estadística descriptiva

Medidas de dispersión

Nos ayudan a determinar de qué forma los valores de cada muestra se alejan del valor medio.



Ejemplo: Comparativa de **dos** máquinas expendedoras de refrescos que no rellenan exactamente los vasos con la cantidad prevista.

Preprocesamiento: Estadística descriptiva

Medidas de dispersión

Error medio

$$\sum_{i=1}^n \frac{|x_i - \bar{x}|}{n}$$

Varianza muestral

$$s^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n - 1}$$

Desviación estándar muestral ó Desviación típica

$$s = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n - 1}}$$

La varianza representa la variabilidad de una serie de datos respecto a su media, también se puede calcular como la desviación típica al cuadrado.

Las máquinas A y B pueden tener la misma media, por ejemplo, $\bar{x}_A = \bar{x}_B = 33$ cl por lo que ambas estarán bien calibradas, pero diferir en su desviación muestral, por ejemplo $s_A = 1$ cl. y $s_B = 0,5$ cl, indicando que en el caso de la máquina A unas veces sus llenados son de 34 cl y otras de 32 cl, mientras que la máquina B su comportamiento es más regular, llenando unas veces 33,5 cl y otras 32,5 cl. La desviación estándar (muestral) nos informa de cuanto alejado (en media) estará una observación x de la media (muestral).

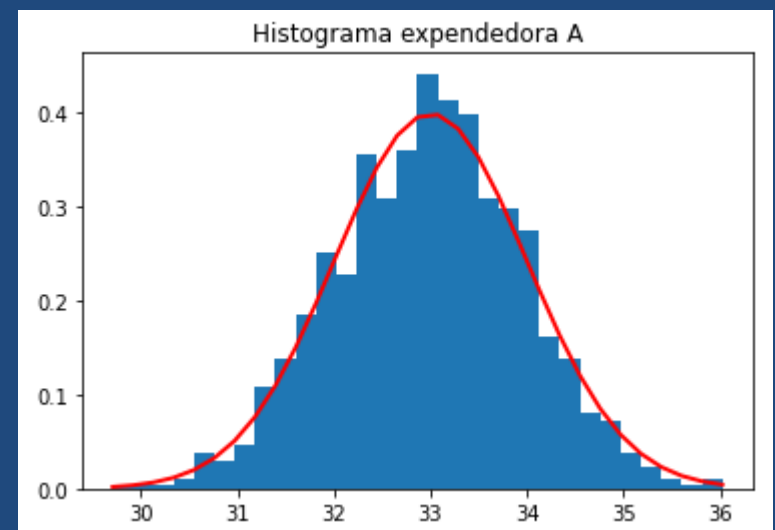
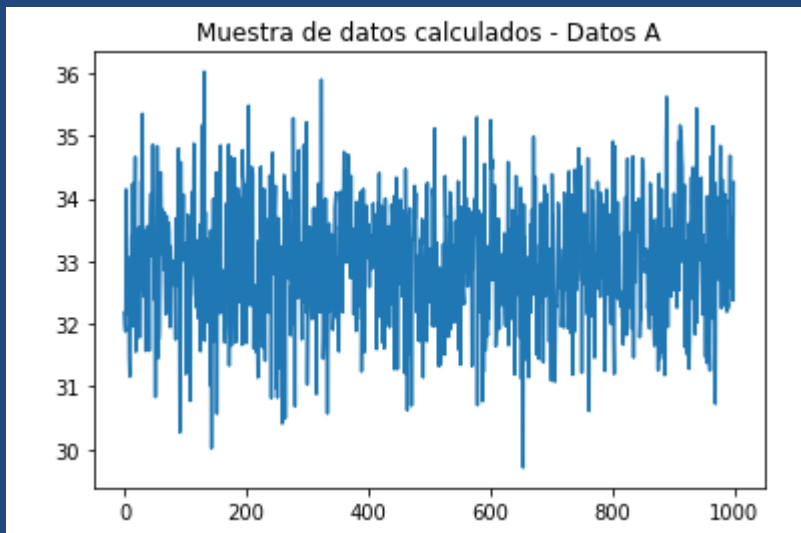
$$\text{Rango} \equiv R = \max_i \{x_i\} - \min_i \{x_i\}$$

Preprocesamiento: Estadística descriptiva

Ejemplo: Medidas de centralidad y dispersión

Título: Ejemplo 2_1: Estadística descriptiva de datos continuos - Expendedora de refrescos

Url: <https://colab.research.google.com/drive/1e-plMYxm6fnLI1sTVEHLFi8xPKGXmFj9?usp=sharing>



Título: Distribución normal

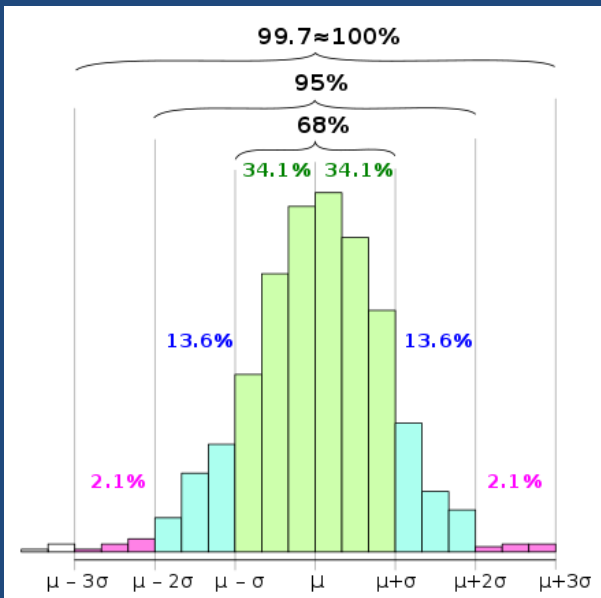
Url: https://es.wikipedia.org/wiki/Distribuci%C3%B3n_normal

Preprocesamiento: Distribución normal

Esta sección está relacionada con el método utilizado para calcular valores aleatorios comprendido en un intervalo. En nuestro caso utilizamos la **distribución normal**.

Por ejemplo:

```
muA, sigmaA = 33, 1 # media y desviación estándar embotelladora A en cl
muB, sigmaB = 33, 0.5 # media y desviación estándar embotelladora B en cl
datosA = np.random.normal(muA, sigmaA, 1000)
datosB = np.random.normal(muB, sigmaB, 1000)
```

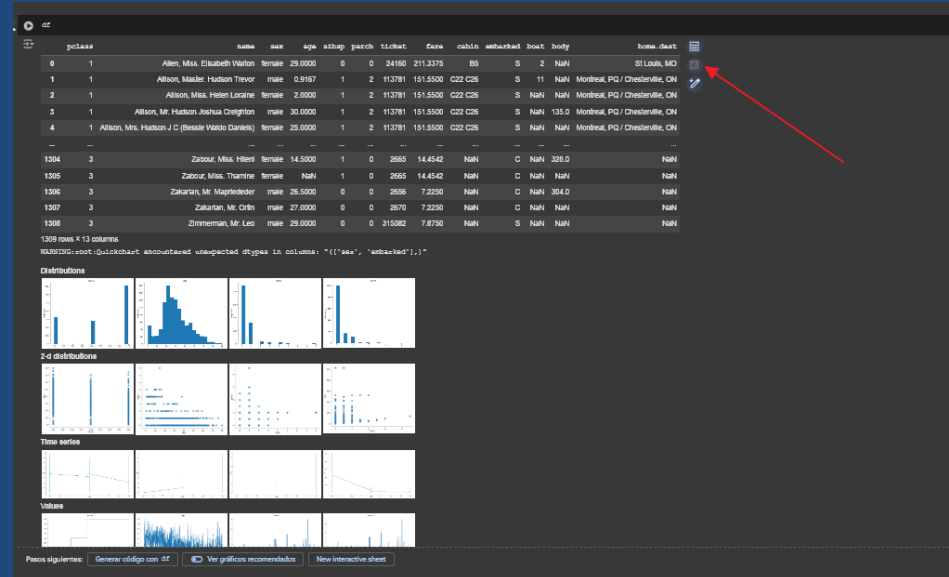
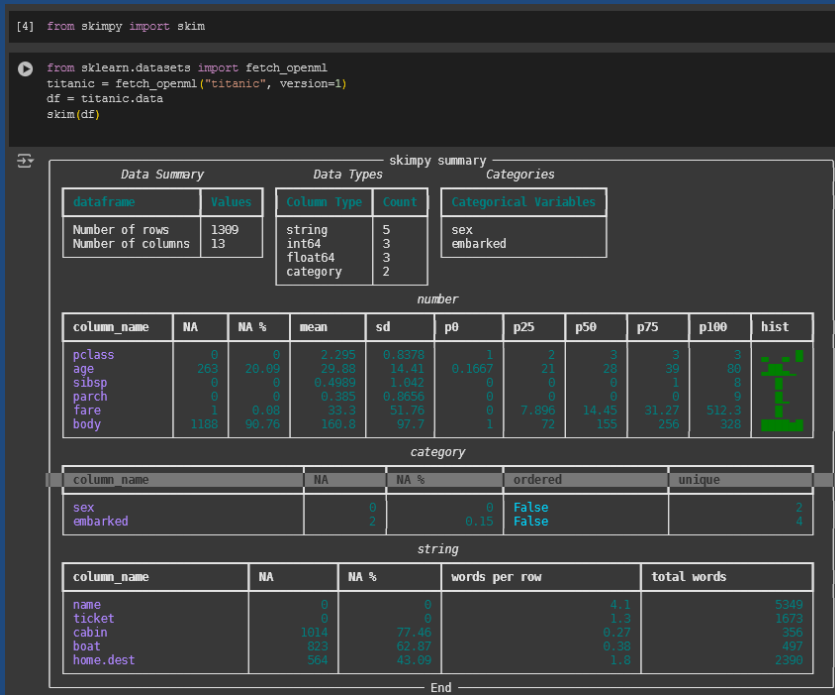


En estadística, la **regla 68-95-99.7**, también conocida como regla empírica, es una abreviatura utilizada para recordar el porcentaje de valores que se encuentran dentro de una banda alrededor de la media. Más exactamente, el 68.27 %, el 95.45 % y el 99.73 % de los valores se encuentran dentro de bandas con semiancho de una, dos y tres veces la desviación típica respecto a la media.

Preprocesamiento: Explorando datos con herramientas

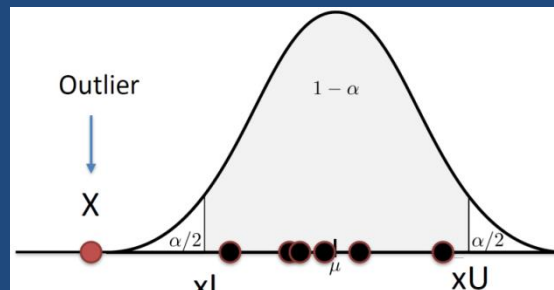
Título: Ejemplo_2_8_Ejemplo uso de la librería Skim - Resumen datos .ipynb

Url: https://colab.research.google.com/drive/1nI9rifxhsex_rI_UD8enLo6g9SICSNaF?usp=sharing



Preprocesamiento: Outliers y observaciones influyentes

Los **outliers** son observaciones que tienen valores inusuales, muy grandes o muy pequeños en relación al resto de la muestra. (El pez grande se como al pez pequeño)



© UCLM

Y por otro lado se encuentran **las observaciones influyentes**: aquellas que impactan en los resultados de un procedimiento estadístico.

A groso modo se podría decir que toda observación influyente es un outlier pero no todo outlier es una observación influyente. Esto último ocurriría si en un análisis cambiásemos el procedimiento de análisis al detectar algún outlier.

Por ejemplo ,si utilizamos la media como procedimiento estadístico para obtener conclusiones, entonces un outlier se convierte en una observación influyente ya que su presencia hace que el valor de la media cambie sustancialmente. Por eso es recomendable eliminarlo. Pero si cambiamos de procedimiento, por ejemplo calculando la mediana, el dato seguiría siendo un outlier pero ya no afectaría a la mediana, y por lo tanto ya no sería influyente.

Preprocesamiento: Outliers y observaciones influyentes

Por ejemplo:

- En un estudio sobre ingresos de la población, que haya una persona con unos ingresos extremadamente altos o bajos en el conjunto de personas estudiadas.

Si un vecino es Amancio Ortega las rentas del pueblo subiría considerablemente haciendo parecer a todos los habitantes ricos. En el cálculo de la mediana la existencia de Amancio no influiría casi nada en el cálculo de la mediana. Amancio seguiría siendo un outlier en el segundo caso pero no una observación influyente para el procedimiento de la mediana.

- Introducción de errores en los datos, por ejemplo un dato con un dígito más respecto al resto. (154 → 3154)

Por lo tanto es necesario/conveniente:

- Detectarlos
- Corregirlos ó extraerlos del estudio o utilizar un método que no sea sensible a estos datos ya que pueden confundir las conclusiones. Por ejemplo que una variable sea significativa cuando no lo sea o al revés
- O tratarlos de forma especial (Ejemplo: diferencias de renta por comunidades durante la pandemia: Canarias y Baleares)

Preprocesamiento: Outliers

El ejemplo **Ejemplo_2_3_Outliers.ipynb** está muy bien porque detecta los outliers.

Url: https://colab.research.google.com/drive/1C6uBUxui_Qq9ee-51ycVYcqigrSHSZNY?usp=sharing

En este ejemplo utiliza el método 1: basado en las bandas (XL y XU). Obtiene los valores de estas bandas en función de una probabilidad (pg) indicada de que los valores/muestras obtenidas estén dentro de dicha banda.

```
Cálculo de bandas

[9] xL= round(np.mean(datos)-Z_alfa* np.std(datos),4)
    xU= round(np.mean(datos)+Z_alfa* np.std(datos),4)
    print(f" Banda= [ {xL},{xU}]")

Banda= [ 1.0044,98.2596]
```

Es decir, todos los valores que están fuera de dichas bandas se consideran outliers.

```
[22] for i in range(len(datos)):
      if datos[i] < xL or datos[i]>xU:
          print(f" El dato[{i}]= {datos[i]} es un outlier")

El dato[50]=100.0 es un outlier
El dato[75]=1.0 es un outlier
```

Preprocesamiento: Outliers

Tras ver el ejemplo anterior, nos planteamos la pregunta **¿Cuándo utilizar un método u otro?**. A continuación se indican una serie de recomendaciones, aunque no son concluyentes dado que dependerá de la naturaleza de los datos y del objetivo del análisis.

Método Basado en la Probabilidad

- Distribución Normal: Este método asume que los datos siguen una distribución normal.
- Datos Científicos y Precisos: A menudo se utiliza en campos como la astronomía o la física, donde se espera que los datos sigan patrones teóricos específicos y los outliers pueden ser debido a errores de medición o anomalías reales.
- Análisis Riguroso: Es útil cuando se requiere un criterio más estricto y matemáticamente definido para identificar outliers.

Método Basado en Cuartiles

- Distribuciones No Normales: Este método no asume una distribución normal.
- Datos de Encuestas y Sociales: Comúnmente usado en estadísticas sociales, económicas y de mercado.
- Análisis Exploratorio: Es una herramienta útil para un análisis exploratorio inicial, proporcionando una forma rápida y fácil de identificar posibles outliers sin necesidad de ajustes complejos.

Preprocesamiento: Observaciones influyentes

Los procedimientos estadísticos, como regresión, clasificación o clustering, pueden estar fuertemente influidos por unas determinadas observaciones-registros (las llamadas observaciones influyentes) que poseen valores extremos en determinadas variables.

Por ejemplo, si estamos ajustando una recta a una nube de puntos, puede existir un punto muy alejado de la nube que fuerza a que la recta de ajuste intente acercarse a él, alejándose del grupo.

Practicar con el ejemplo Ejemplo_2_4_Observaciones_influyentes_Sin_soluciones.ipynb, está dedicado a implementar los procedimientos para determinar observaciones influyentes en el cálculo de la media y en la mediana.

Url: https://colab.research.google.com/drive/11JM5daNQUCB_VSAOHmpFjHmFuZDsB3-i?usp=sharing

Comentarlo y hacer pruebas, proponer los ejercicios de la actividad:

Actividad 2.1 – Análisis de observaciones influyentes

Preprocesamiento: Escalamiento (Estandarización)

El **orden de magnitud** de las variables influyen en los procedimientos estadísticos.

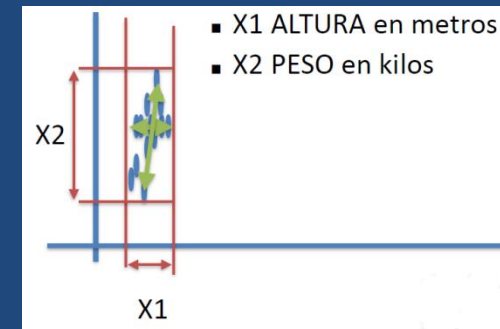
Ejemplo de análisis estadístico de personas:

Altura en metros \rightarrow intervalo $[0,2]$ metros

Peso en kilos \rightarrow intervalo $[0,200]$ Kgs



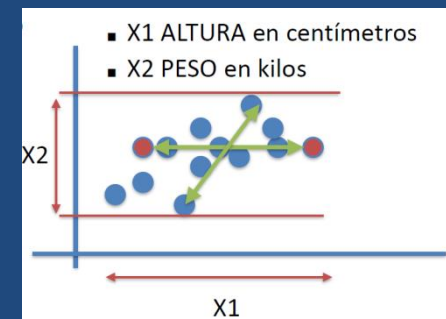
El peso tiene un rango 100 veces superior a la altura



© UCLM

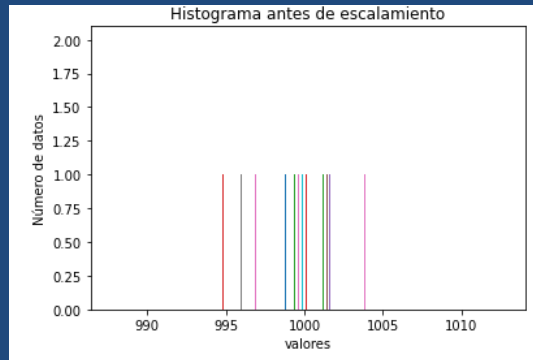
Si fuésemos a calcular una distancia entre individuos, la componente relativa a la altura sería irrelevante frente al peso

El objetivo es que ninguna variable esté dominada por otra.



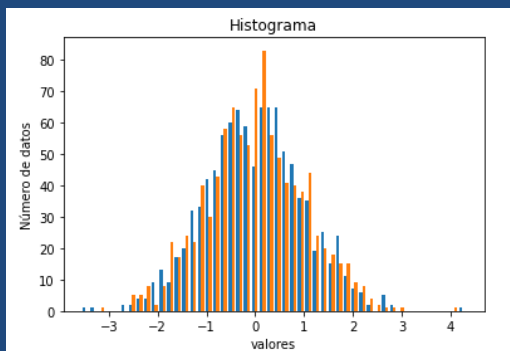
© UCLM

Preprocesamiento: Escalamiento (Estandarización)



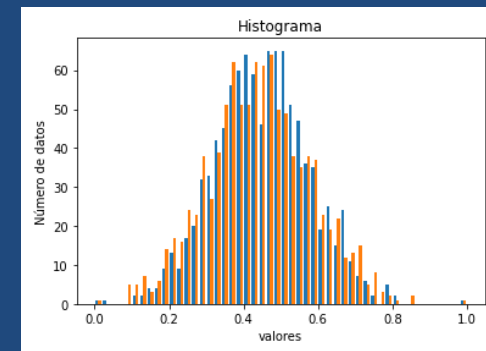
Estandarización por rangos
Se reemplazan las variables por:

$$x_i^{new} = \frac{x_i^{old} - \min(x_i^{old})}{\max(x_i^{old}) - \min(x_i^{old})}$$



Estandarización Z – score
Se reemplazan las variables por:

$$x_i^{new} = \frac{x_i^{old} - \bar{x}_i}{s_i}$$



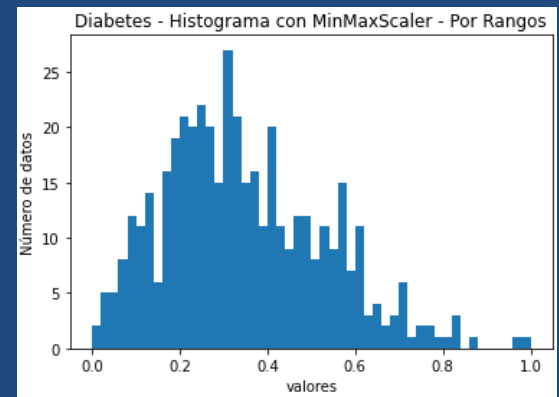
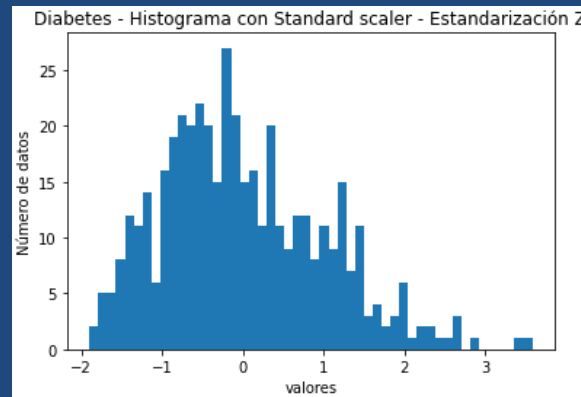
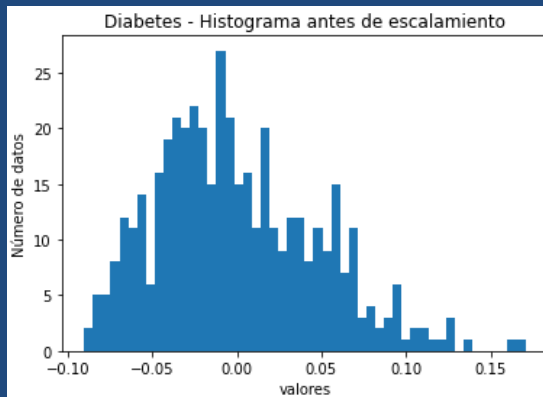
**Método
más
efectivo**

Preprocesamiento: Escalamiento (Estandarización)

Veamos los siguientes ejemplos de escalamiento con dos datasets reales:

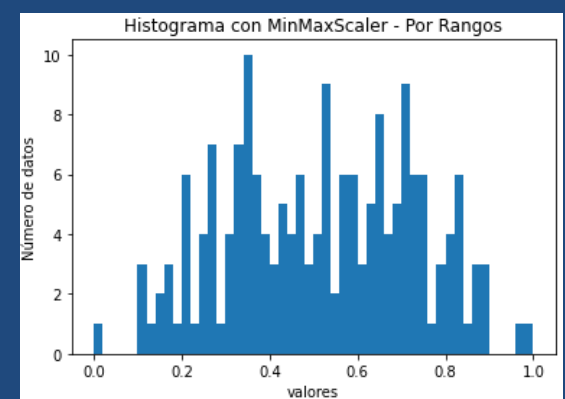
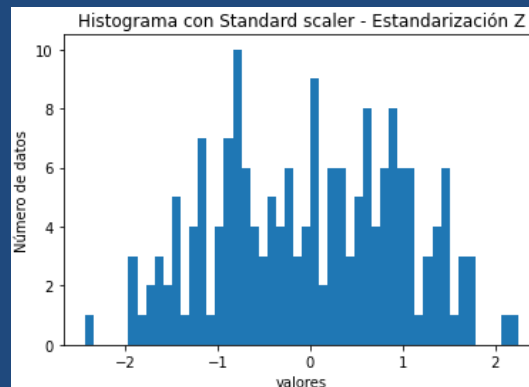
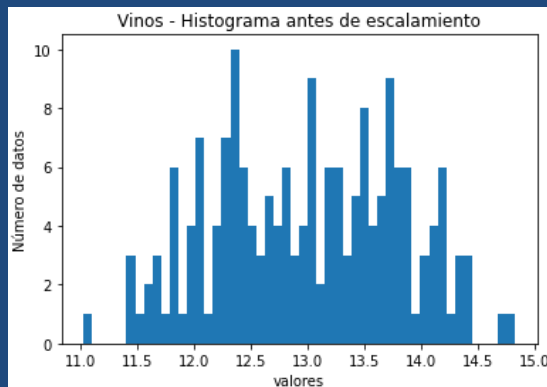
Ejemplo_2_6_Escalamiento_de_datos_Diabetes.ipynb

Url: <https://colab.research.google.com/drive/1bY51scSkFpVYYL8LS3utFS1PCQ2gHAX7?usp=sharing>



Ejemplo_2_7_Escalamiento_de_datos_Vinos.ipynb

Url: <https://colab.research.google.com/drive/11sID8FJqBjwnTb9rylOTnJQ9mNVnpJx3?usp=sharing>



Observamos que la distribución de los datos es la misma después del escalamiento

Preprocesamiento: Selección de variables y Ponderación de variables

- Las bases de datos pueden contener miles de variables, causando problemas de computación e interpretación. Un primer problema que hay que abordar es determinar qué conjunto de estas variables serán empleadas en el análisis.
- En el **análisis cluster** se suele recurrir a la reducción de la dimensionalidad de la matriz de datos X mediante análisis de componentes principales que transforma la anterior matriz a una matriz X' de dimensiones mucho menor que la original. (normalmente se reduce el número de columnas)
- En problemas de **clasificación y regresión** se utilizan las medidas de correlación entre la variable respuesta y el conjunto de regresores (características) x_i quedándose con las que tienen mayores correlaciones.
- Un analista de datos se enfrenta al siguiente dilema: Emplear todos los datos para aprovechar así toda la información y obtener así modelos más precisos o emplear exclusivamente las variables relevantes. (**Principio de parsimonia**)

Preprocesamiento: Selección de variables y Ponderación de variables

- El problema anterior aborda cómo elegir un subconjunto significativo de variables. Una vez resuelto, todavía subyace el problema de que no todas las variables son igualmente importantes. Cuando hemos estandarizado las variables y las hemos homogeneizado para poderlas comparar.
- Inconvenientes de utilizar un conjunto de variables muy alto:
 - Incremento del coste computacional.
 - En su interpretación: No se llega a conocer cuáles son el conjunto de variables esenciales.
 - Se va a aplicar un modelo de regresión/clasificación, entonces una vez entrenado si se va a analizar un nuevo caso, entonces se hace necesario obtener todas las variables que se utilizaron en el entrenamiento (ejemplo: devolución de préstamo y color del coche)
- En análisis cluster una forma de resolver este problema es asignar pesos a las variables en el cálculo de distancia. En problemas de regresión/clasificación son los propios métodos quienes llegan a ponderar implícitamente las variables estandarizadas.



Preprocesamiento de datos con Excel

(Hay muchos recursos en internet)

Título: Video - Limpieza de hojas de calculo "Data cleaning" de #excel #excel365

Url: https://www.youtube.com/watch?app=desktop&v=kh3sDmX21_k

Título: 13 Steps for Data Cleaning in Excel

Url: <https://www.linkedin.com/pulse/13-steps-data-cleaning-excel-william-irvin-et4sc/>

Título: Data Cleaning: Beginner Excel Guide

Url:

https://www.linkedin.com/feed/update/urn:li:activity:7198510140474458112?utm_source=share&utm_medium=member_desktop



Preprocesamiento: Datasets no balanceados / desbalanceados

- En el ejemplo siguiente se trabaja con un Dataset de datos no balanceados y creo que poder ser un buen ejemplo sobre el que desarrollar este tema:

Título: Ejemplo_3_9_Árboles_de_Decisión_para_Clasificación_(trading)_v2.ipynb

(Lo veremos en la próxima UT3)

- **Título:** Business Intelligence, Analytics & Data Visualization (Moderated).
Explica lo que son los Dataset no balanceados, explica los casos donde suele ocurrir y una técnica para resolverlo.
- **Url:** https://www.linkedin.com/feed/update/urn:li:activity:7170148606996377601?utm_source=share&utm_medium=member_desktop
- **Título:** Qué son los Datos Desbalanceados y Cómo balancearlos usando Submuestreo y Sobremuestreo con Python
- **Url:** <https://www.youtube.com/watch?v=2J90FG6QKL4>

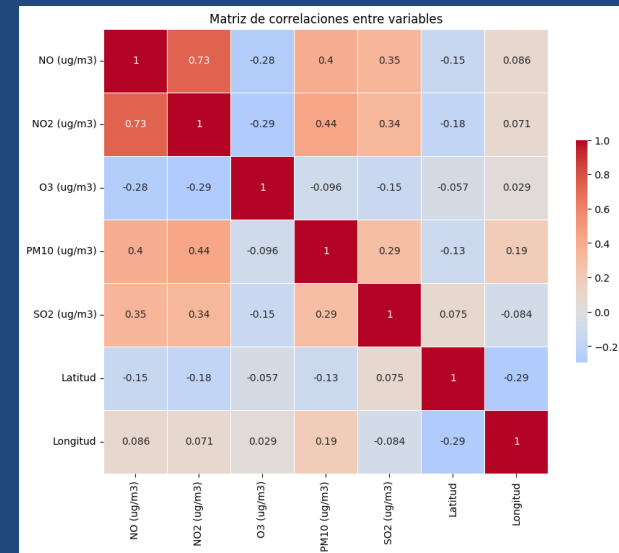
Guía práctica de introducción al análisis exploratorio de datos en Python



datos.gob.es
reutiliza la información pública

Título: Guía práctica de introducción al análisis exploratorio de datos en Python

Url: https://datos.gob.es/es/documentacion/guia-practica-de-introduccion-al-analisis-exploratorio-de-datos-en-python?utm_source=newsletter&utm_medium=email&utm_campaign=Gua-para-el-analisis-exploratorio-de-datos-en-python-nuevo-podcast-sobre-datos-de-alto-valor-y-mucho-ms-en-datosgobes



Contiene un ejemplo de herramienta en la última celda que genera una página HTML con el análisis exploratorio realizado

Cuaderno demo: Panda, Numpy, Matplotlib y carga de ficheros

Proponer la actividad Actividad 2.2 – Cuaderno demo UT2 - Ejercicios de ampliación

Url: <https://colab.research.google.com/drive/1XPKvv3BuHdd8GUPNiTsjd50M8HBN4Lv-?usp=sharing>



```
# CARGAMOS LIBRERIAS
# =====
# Importamos la librería NUMPY para CÁLCULO NUMÉRICO Y ANÁLISIS DE DATOS
import numpy as np
# PANDAS para MANIPULACIÓN y ANÁLISIS DE DATOS
import pandas as pd
# MATPLOTLIB para GRAFICOS
import matplotlib.pyplot as plt
# SCIPY para CÁLCULO MATEMÁTICO
import scipy.stats as st
# Preprocesado
# =====
from sklearn.compose import ColumnTransformer
```