



Posgrado Smart Energy

DIGITAL ENERGY

Sesión 3 – Estadística Descriptiva y Análisis Exploratorio de Datos (EDA)

Marc Jené: marc.jene@upc.edu

Sara Barja: sara.barja@upc.edu



Calendario

	Lunes	Martes	Miércoles	Jueves
ABRIL	18	19	20 S1 – Introducción a Machine Learning	21
	25 S2 – Introducción a Python	26	27 S3 – Estadística descriptiva	28
MAYO	2 S4 – Modelos de aprendizaje supervisado (I): Clasificación	3	4 S5 – Modelos de aprendizaje supervisado (II): Regresión	5
	9 S6 – Aplicación de AI en el sector eléctrico: Odit-e	10	11 S7 – Modelos de aprendizaje no supervisado	12
	16 S8 – Ejercicio final			

● Entrega Práctica



Objetivos del módulo

- Diferenciar muestra de población.
- Aprender a limpiar datasets y obtener resumen de datos mediante la librería pandas.
- Diferenciar tipos de variables.
- Conocer parámetros básicos de estadística descriptiva: la media, la mediana, la moda, la variabilidad...
- Introducción a la variabilidad.



Análisis Exploratorio de Datos

- ¿Qué es el análisis exploratorio de datos?

El **EDA** (Exploratory Data Analysis) es una forma de analizar datos, un proceso crítico que consiste en realizar **investigaciones iniciales** sobre los datos para descubrir patrones, detectar anomalías, probar hipótesis y comprobar supuestos con la ayuda de estadísticas de resumen y representaciones gráficas.

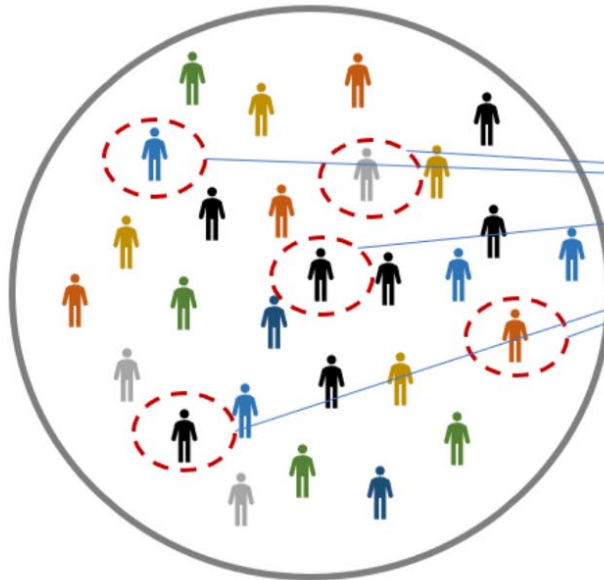
Es una buena práctica **entender primero los datos** y tratar de obtener la mayor cantidad de información posible de ellos. El EDA consiste en **dar sentido a los datos que se tienen a mano, antes de empezar a tratarlos y ensuciarlos**.



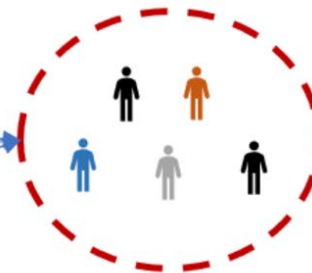
Muestra vs. Población

Normalmente nos encontraremos en una situación donde deseamos responder preguntas sobre una *población* pero solo tenemos acceso a una *muestra*.

Para responder
preguntas sobre una
Población



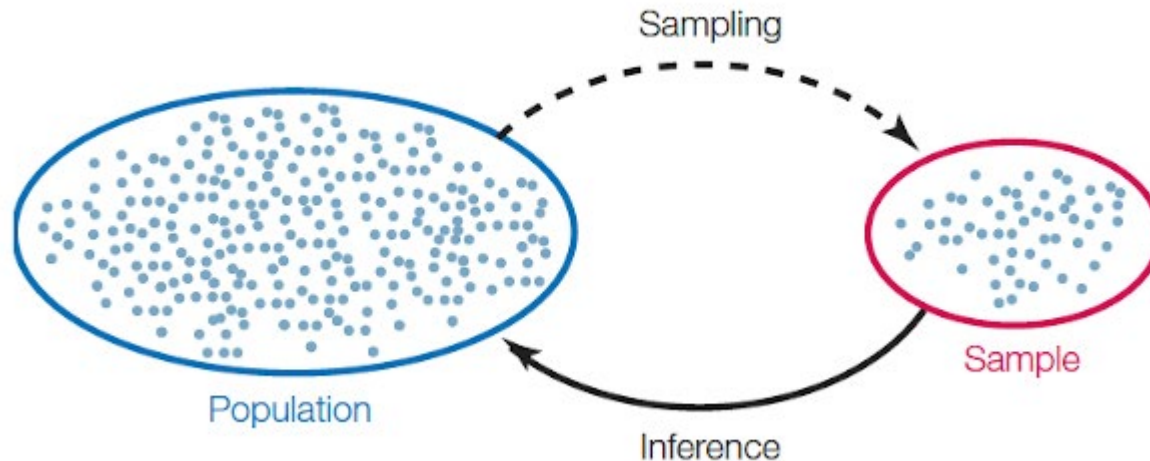
Tenemos que aprender
a trabajar con una
Muestra





Muestra vs. Población

La *población* se refiere a todos los individuos que son relevantes para una pregunta en concreto, mientras que una *muestra* será solo un subset de estos. Por ejemplo, todos los clientes de una compañía de distribución serán la población, mientras que para hacer un estudio a lo mejor solo usamos una muestra de ellos. A veces lo que para una pregunta es una *población*, para otra será una muestra (en nuestro ejemplo todos los clientes de una distribuidora son solo una muestra de la población de un país).

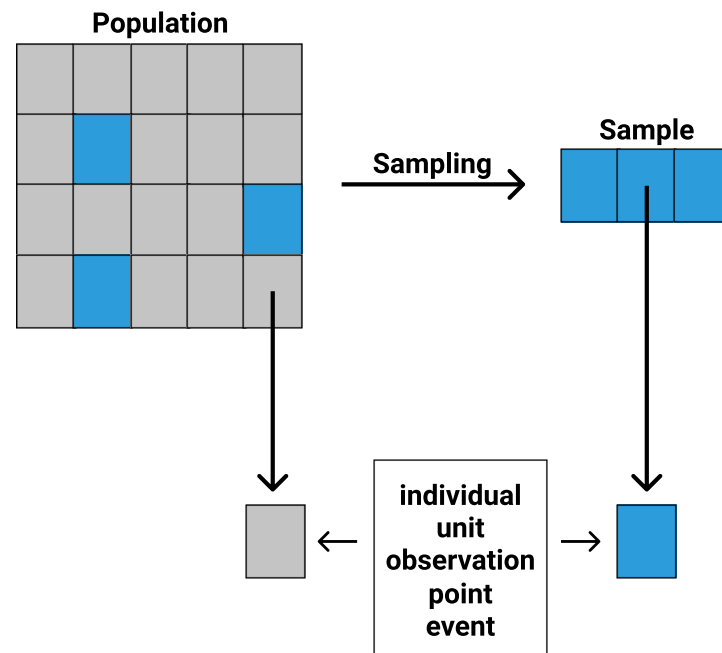




Muestra vs. Población

Las poblaciones y las muestras están formadas por varias observaciones, individuos, elementos, etc.

Siempre que se pueda, será mejor usar la población para responder a nuestras preguntas, pero a veces esto no es posible (no se tienen todos los datos, no todos los clientes tienen un Smart meter, recopilar todos los datos en una misma fuente es difícil, etc.). En estos casos usaremos una muestra.





Muestra vs. Población - Resumen

Población: todos los individuos que son relevantes para una pregunta en concreto.

Muestra: subset de la población.

Muestreo: obtener muestra de población.

Inferencia: obtener conclusiones para una población partiendo de una muestra.

$$\text{error de estimación} = \text{parámetro (población)} - \text{estadística (muestra)}$$

Para mejorar el muestreo:

- Varias muestras y trabajar con promedios.
- Intentar acercarnos al máximo a toda la población (más observaciones).
- Muestreo estratificado (*stratified sampling*), para que sea significativo.



Limpieza y resumen de datos

Antes de ajustar un modelo de ML, **siempre tenemos que limpiar los datos**. Ningún análisis crea resultados significativos con datos confusos.

La limpieza de datos (*data cleaning*) o depuración de datos (*data cleansing*) es el proceso de detectar y corregir (o eliminar) los datos corruptos o inexactos de un dataset.

Consta de dos procesos básicos:

- Identificación de las partes incompletas, incorrectas, inexactas o irrelevantes de los datos.
- Sustitución, modificación o eliminación de los datos sucios o inadecuados.

A la práctica, los datos brutos son casi siempre desordenados. Si utilizas esos datos para el análisis, por ejemplo, alimentando un modelo de aprendizaje automático, obtendrás conocimientos inútiles o erróneos la mayoría de las veces.

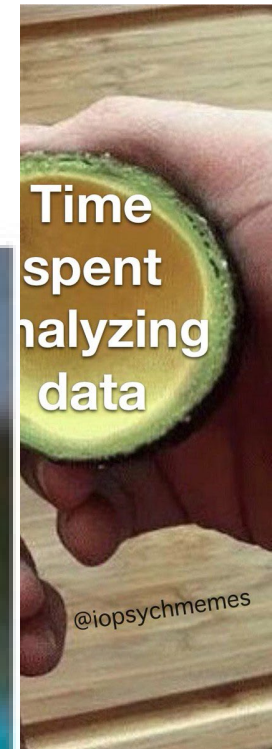


Limpeza y resumen de datos

Desde luego, la limpieza de datos no es divertida y requiere mucho tiempo.



**When you find out Machine Learning
really means endless data cleaning**





Limpieza y resumen de datos

Un *data scientist* pasa el 80% de su tiempo en el trabajo limpiando datos desordenados en lugar de hacer un análisis de estos o preparar modelos de inteligencia artificial.

La librería *pandas* nos ofrece varias opciones para obtener un resumen de los datos como el método *describe*, u otros como *sum*, *count*, *min*, *max*, *mean*... A parte, también son interesantes para obtener distribuciones de frecuencia los métodos *value_counts* i *nunique*.



Tweet



Big Data Borat @BigDataBorat · Feb 27, 2013

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.





Limpieza y resumen de datos

Hay varias opciones para tratar con valores vacíos, pero pandas nos ofrece algunas opciones rápidas e interesantes para ir rápido.

Funciones interesantes

df.isna() – Detecta los valores ausentes/nulos (NA).

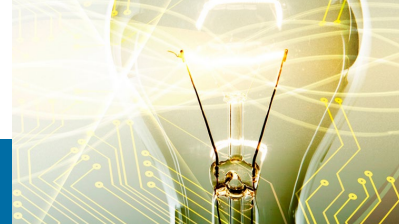
df.dropna() – Elimina las filas con cualquier columna que tenga datos NA.

df.fillna() - Sustituye todos los datos NA por el valor deseado. Este método con los valores NA de tres maneras distintas.

- `method = 'bfill'`: propaga el primer valor no nulo observado hacia atrás hasta que se encuentra otro valor no nulo
- `method = 'ffill'`: propaga el último valor no nulo observado hasta que se encuentra otro valor no nulo
- `method = explicit value`.

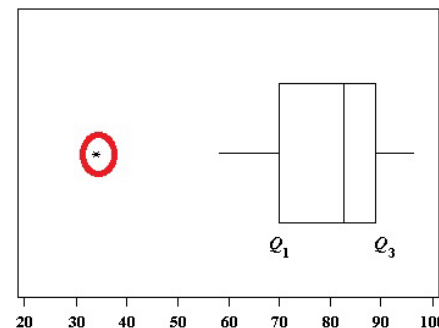
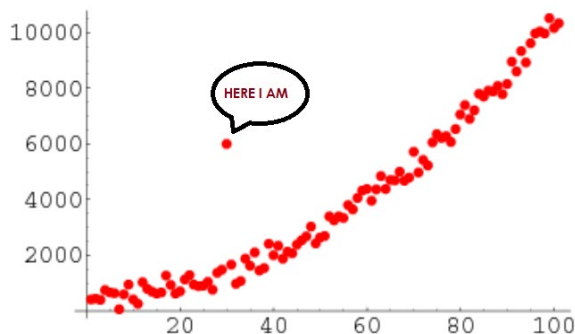
También nos puede interesar el método **df.interpolate()**.





Limpieza y resumen de datos – Pasos a seguir

- **Muestreo de datos.** Si hay muchos datos y el tiempo de cómputo puede ser demasiado grande se trabaja primero con solo una muestra
- **Valores ausentes.** Es usual que muchos parámetros estén vacíos. Para pasar a pasos posteriores esto se debe arreglar. Si faltan valores, podemos:
 - Eliminar toda la fila.
 - Inferir otro valor.
- **Valores atípicos (outliers).** Son valores inusuales en un conjunto de datos. Los valores atípicos son problemáticos para muchos análisis estadísticos porque pueden hacer que las pruebas no detecten resultados significativos o distorsionen los resultados reales. Dependiendo de cada caso, se deben considerar y tener en cuenta o pueden tratarse igual que los valores ausentes (se pueden eliminar o sustituir por valores típicos).





Limpieza y resumen de datos – Sigüientes pasos

- **Normalizar.** Transformar los datos a distribuciones normalizadas.
- **Reducir dimensiones.** Muchas variables nos pueden causar problemas en Machine Learning. Para eliminar variables podemos:
 - a) Eliminar variables irrelevantes.
 - b) Eliminar variables redundantes.
- **Añadir dimensiones.** En otros casos haremos lo contrario y transformaremos o añadiremos variables (convertir un código postal coordenadas en unas coordenadas, crear grupos, etc.).
- **Discretizar variables numéricas en categorías.** Pasar de una variable continua a categórica (edad a tramos de edad).
- **Binarizar categorías.** En algunos casos de machine learning, tendremos que usar variables binarias.

Human-Readable

Machine-Readable

Pet	Cat	Dog	Turtle	Fish
Cat	1	0	0	0
Dog	0	1	0	0
Turtle	0	0	1	0
Fish	0	0	0	1
Cat	1	0	0	0



Estadística descriptiva

Tipos de variables

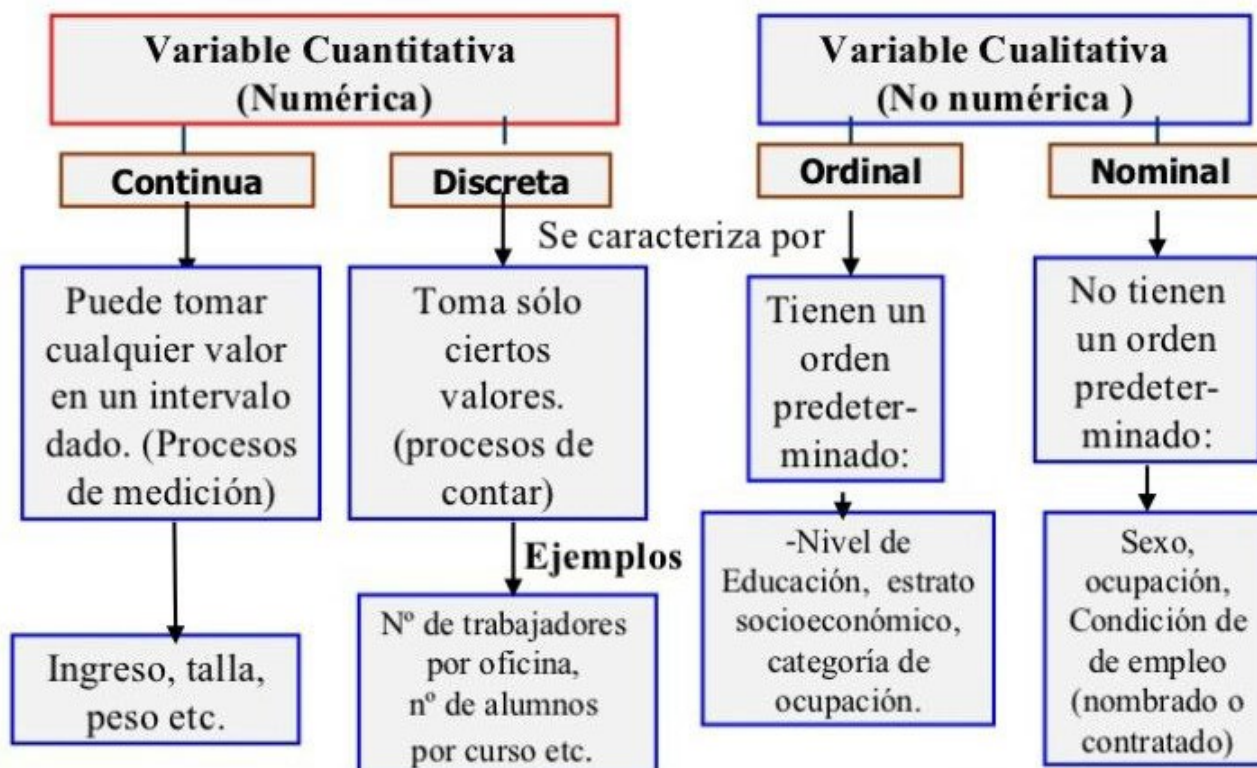
Las variables pueden ser nominales, ordinales, continuas o discretas.

Tipo de variable	Cero verdadero	Intervalos iguales	Orden	Categoría	Ejemplo
Nominal	No	No	No	Sí	Estado civil, género, nacionalidad
Ordinal	No	No	Sí	Sí	Estrato socioeconómico, nivel educación, satisfacción
Discreta	No	Sí	Sí	Sí	Años, número de hijos, páginas de un libro
Continua	Sí	Sí	Sí	Sí	Altura, peso, consumo mensual, temperatura (°C)



Estadística descriptiva

Clasificación de Variables



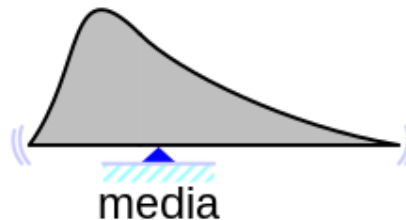
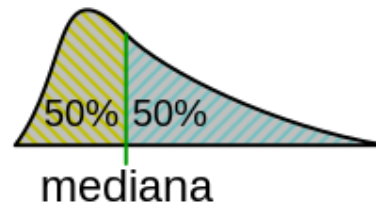
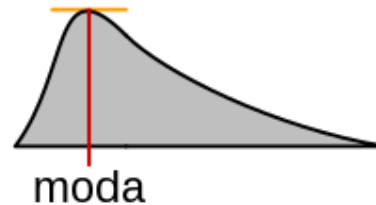


Estadística descriptiva

Media, mediana y moda

Una vez tenemos un resumen de los datos usados, unos parámetros que nos pueden ser muy útiles para conocer como se distribuyen ciertas características de nuestro *dataset* son:

- Media aritmética
 - Mediana
 - Moda
- + Media ponderada



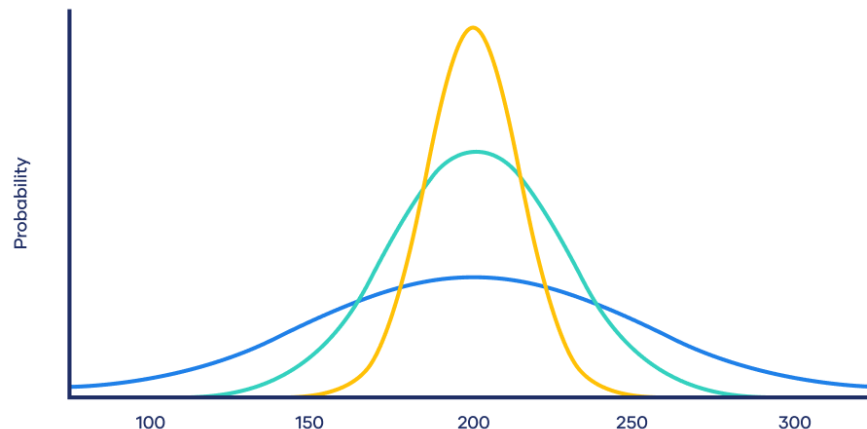


Estadística descriptiva

Variabilidad

La variabilidad es una medida de la dispersión de los datos en una distribución, sea esta teórica o de una muestra; medidas de variabilidad son

- la varianza
- la desviación estándar
- cuartiles o deciles
- rango.





Estadística descriptiva

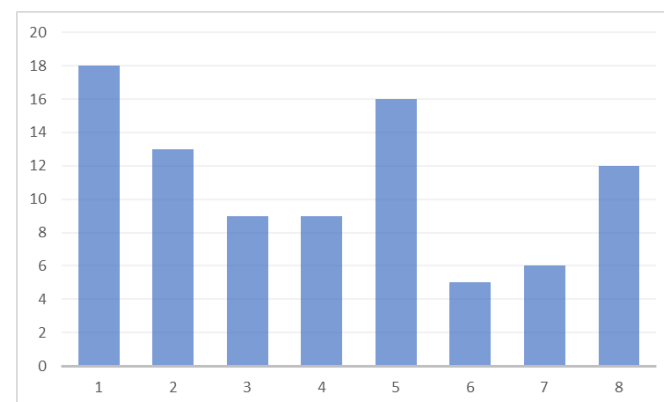
Tablas de frecuencia

Las tabla de frecuencias muestran de forma ordenada un conjunto de datos estadísticos y a cada uno de ellos le asigna una frecuencia (las veces que se repite un número o dato). Hay distintos tipos de frecuencias:

- Frecuencia absoluta
- Frecuencia absoluta acumulada
- Frecuencia relativa
- Frecuencia relativa acumulada

La mejor manera de comparar y evaluar distintas tablas de frecuencia es visualizarlas.

Xi	Frecuencia absoluta (ni)	Frecuencia absoluta acumulada (Ni)	Frecuencia relativa (fi=ni/N)	Frecuencia relativa acumulada (Fi=Ni/N)
1	18	18	0.20	0.20
2	13	31	0.15	0.35
3	9	40	0.10	0.45
4	9	49	0.10	0.56
5	16	65	0.18	0.74
6	5	70	0.06	0.80
7	6	76	0.07	0.86
8	12	88	0.14	1.00
Total	88		1	

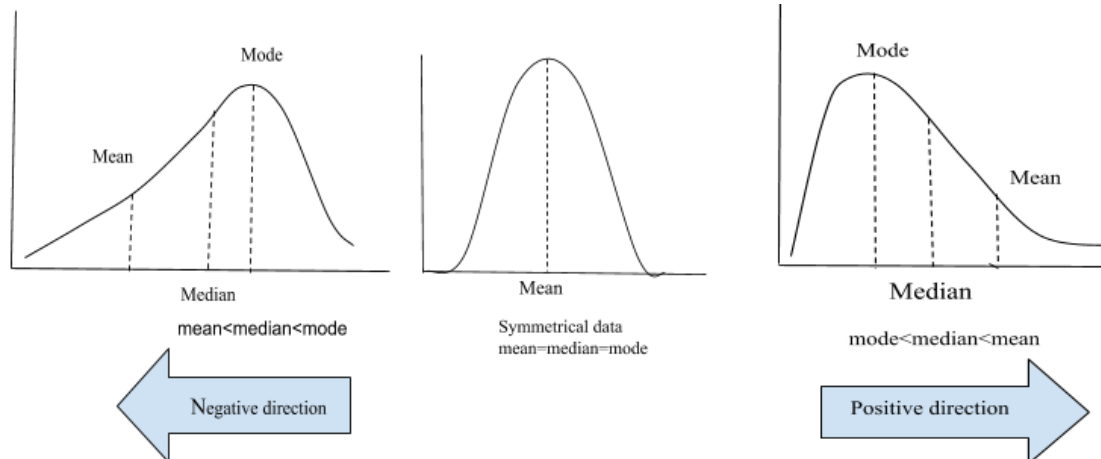




Estadística descriptiva

Asimetría (*skewness*)

Skewness se refiere a una distorsión o asimetría que se desvía de la curva de campana simétrica, o distribución normal, en un conjunto de datos. Si la curva se desplaza hacia la izquierda o hacia la derecha, se dice que está sesgada.



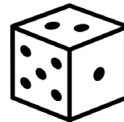


Probabilidad

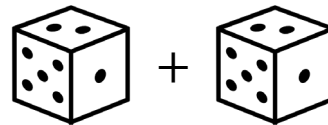
Aunque hay muchas veces que somos capaces de predecir el resultado de una operación, en muchas ocasiones solo podremos trabajar con probabilidades. Por ejemplo, para una moneda tenemos:



$$P(\text{cruz}) = 0.5$$



$$P(\text{Número } 2) = ?$$



$$P(2 \text{ i } 2) = ?$$





Probabilidad

Hay dos normas básicas para calcular probabilidades combinado dos posibles eventos.

$$P(A \text{ y } B) = P(A) \cdot P(B)$$

$$P(A \text{ o } B) = P(A) + P(B) - P(A) \cdot P(B)$$

O introduciendo una nueva notación:

$$P(A \cup B) = P(A) + P(B) - P(A) \cap P(B)$$

Otra norma importante es:

$$P(A \cup \text{no } A) = P(A) + P(\text{no } A) = 1$$



Probabilidad

Eventos independientes

Cuando los eventos son independientes es cuando podemos usar la fórmula

$$P(A \text{ y } B) = P(A) \cdot P(B)$$

Por ejemplo, la probabilidad de sacar 4 veces una cruz:

$$P(\text{Cruz 4 veces}) = P(\text{Cruz}) \cdot P(\text{Cruz}) \cdot P(\text{Cruz}) \cdot P(\text{Cruz})$$

Pero si los eventos no son independientes, no podemos usarla directamente. Por ejemplo:

$$P(A) = P(\text{Número par}) = \frac{3}{6}$$

$$P(A \cap B) = ??$$

$$P(B) = P(\text{Número menor a 4}) = \frac{3}{6}$$



Probabilidad

Combinatorias y permutaciones

¿Importa el ORDEN?	¿Usamos TODOS	¿Se pueden REPETIR?		EJEMPLO
NO		NO	$C_{m,n} = \binom{m}{n}$	(Mezclar 2 colores diferentes con las 10 pinturas que tengo)
		SÍ	$CR_{m,n} = \binom{m+n-1}{n}$	(Comprar 4 pasteles en una tienda donde tienen 6 tipos de ellos)
SÍ	NO	NO	$V_{m,n} = m \cdot (m-1) \cdot (m-2) \cdot \dots \cdot (m-n+1)$	(Dar 2 premios diferentes entre los 25 alumnos de una clase)
		SÍ	$VR_{m,n} = m^n$	(Números de 3 cifras se pueden formar con los dígitos 2, 4, 6, 8 y 0)
	SÍ	NO	$P_n = n !$	(Formas de sentarse 5 personas en un banco donde caben los cinco)
		SÍ	$PR_n^{n_1, n_2, \dots, n_r} = \frac{n !}{n_1 ! \cdot n_2 ! \cdot \dots \cdot n_r !}$	(Con las cifras 2, 3, 3, 4, 4 y 4 ¿cuántos números de 6 cifras puedo escribir?)



Entregable

Clasificación de usuarios a través del consumo eléctrico de contadores.

Objetivo: Desarrollar un modelo de ML de tipo Supervisado para **Clasificar** a los usuarios de una Comercializadora, según su perfil de consumo eléctrico horario durante un día. Dos tipos de perfiles de cliente: usuarios con un perfil de consumo **alto** y usuarios con un perfil de consumo **no alto**.

A continuación, se detallan algunas pautas generales para desarrollar el modelo.

- 1) Entender el conjunto de datos. Utilice estadísticas descriptivas y herramientas de visualización de datos.
 - 2) Preparar el conjunto de datos. Limpiar y transformarlo, si fuera necesario.
 - 3) Pensar en la posible selección y creación de nuevas características para su modelo. Podría ser, por ejemplo, la media de consumo diario.
 - 4) Dividir el conjunto de datos en entrenamiento y prueba (y validación).
- + ...



Entregable

Clasificación de usuarios a través del consumo eléctrico de contadores.





Bibliografía recomendada

1. EDA <https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>
2. EDA y limpieza de datos
<https://towardsdatascience.com/exploratory-data-analysis-and-data-cleaning-practical-workout-2a20442b42fb>
3. Pipeline de EDA y limpieza de datos
<https://medium.com/@oluwabukunmige/pipeline-for-exploratory-data-analysis-and-data-cleaning-6adce7ac0594>
4. Funciones pandas
<https://pandas.pydata.org/docs/reference/frame.html>