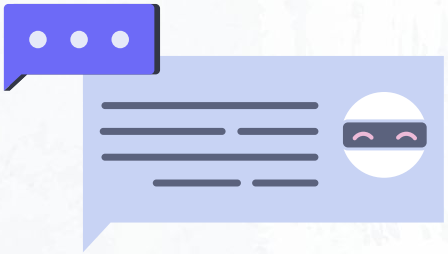
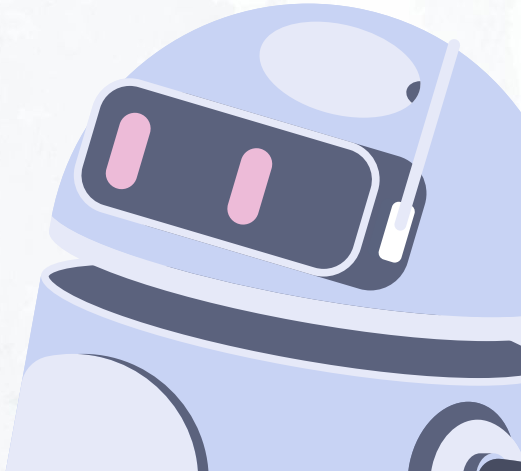
Analítica de Datos (Aprendizaje de máquina)





UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD NACIONAL DE COLOMBIA

Enfoque de análisis

Enfoque supervisado

En el enfoque supervisado se tienen etiquetas para los datos, es decir, existe una columna en donde se define una categoría o clase que identifique a un grupo de datos. Por ejemplo, en una base de datos de vibraciones mecánicas, la etiqueta de clase puede tomar valores

{buen_estado , mal_estado}

Enfoque no supervisado

A diferencia del enfoque anterior, en el enfoque no supervisado no se tienen etiquetas de los datos, es decir, no se conoce la clase a la que pertenecen los datos en la base de datos.

Datos faltantes

Missing values

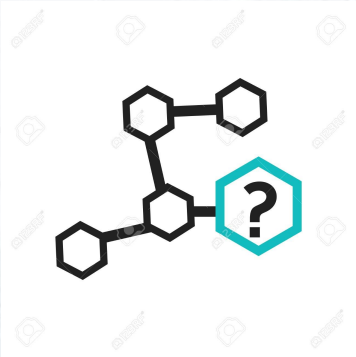
También conocidos como **missing values**, son aquellos valores dentro de las características que por alguna razón (falla en el instrumento de medición, omisión, o cualquier otra) este no se conoce.

Row No.	class	duration	wage-inc-1st	wage-inc-2nd	wage-inc-3rd
1	good	1	5	?	?
2	good	2	4.500	5.800	?
3	good	?	?	?	?
4	good	3	3.700	4	5
5	good	3	4.500	4.500	5
6	good	2	2	2.500	?
7	good	3	4	5	5
8	good	3	6.900	4.800	2.300
9	good	2	3	7	?
10	good	1	5.700	?	?
11	good	3	3.500	4	4.600
12	good	2	6.400	6.400	?
13	bad	2	3.500	4	?

En **Python**, más específicamente con la librería de Pandas, estos valores faltantes se suelen identificar por las variables: **None**, **NaN**, **NA**

Datos faltantes

¿Qué hacer?



- **Eliminar aquellas filas que presenten valores faltantes.**
  Pros
  - El futuro análisis no caerá en errores por información ficticia.

  Contras
  - Pérdida de valiosos datos.
  - Peligro de quedarse con una base de datos muy pequeña y poco representativa.

- **Imputar valores faltantes (variables continuas/categóricas)**
  Pros:
  - Prevención de pérdida de datos.

  Contras:
  - Puede causar fuga de información (data leakage)

Table 1

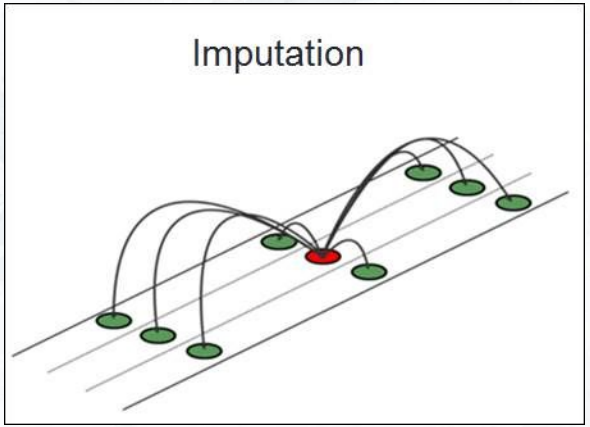
	x1	x2
1	NA	1
2	2	2
3	3	3
4	NA	4
5	2	5
6	3	6

Datos faltantes

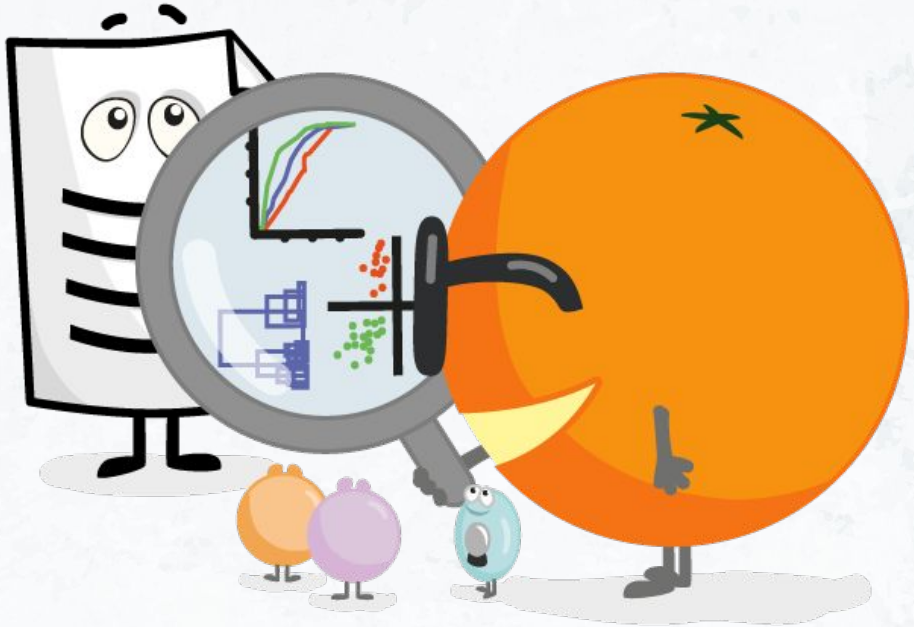
Imputación de datos

Existen varios enfoques para la imputación de datos , entre ellos los principales son:

- **Calcular la media/mediana del atributo** (característica) con los valores conocidos (no missing values) y reemplazar todos los missing values de dicha característica por el resultado. Se aplica a variables numéricas.
- **Calcular la moda o valor con mayor frecuencia del atributo**, y reemplazar los missing values por este. Se aplica a variables categóricas.
- **Usar un método de clasificación** (comúnmente se usa árboles de decisión) para la predicción del valor faltante. Se aplica para variables tanto numéricas como categóricas.



Datos faltantes



Variables numéricas - Transformación

De numericas a categóricas

También conocido como binning o discretización. Consiste en agrupar los valores continuos en rangos de valores.

Ejemplo: discretización de la edad en rangos: [18-24], [25-35], [36-60],[61,~]

Pros:

- Ayuda a mejorar la precisión de los modelos predictivos, debido a que ayuda a reducir el ruido de los datos y la no-linealidad.
- También permite una fácil identificación de outliers o valores de rango.

Esta puede ser supervisada o no-supervisada

Variables numéricas - Transformación

Discretización no-supervisada

En este enfoque no se hace uso de la etiqueta o información de la clase objetivo. Se divide en:

- Discretización de ancho uniforme.
- Discretización de frecuencia uniforme.

Discretización de ancho uniforme.

Los datos se dividen en *k* intervalos del mismo ancho

$$w = \frac{max - min}{k}$$

Se ajustan los límites de los intervalos como:

$$min + w, min + 2w, \ldots, min + (k-1)w$$

Variables numéricas - Transformación

Discretización no-supervisada

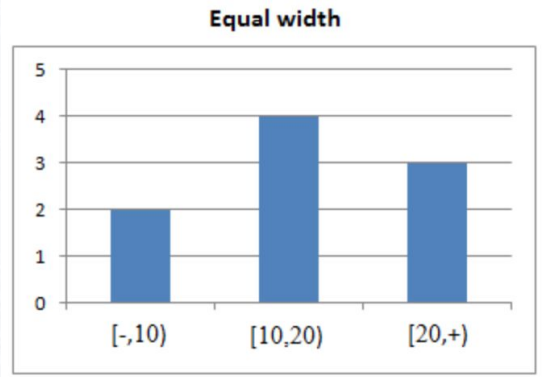
Discretización de ancho uniforme. EJEMPLO:

Sea X el vector de datos $\{X = 0,4,12,16,16,18,24,26,28\}$ y se quiere dividir en **3 intervalos**, k=3.

$$w = \frac{max - min}{k} = \frac{28 - 0}{3} = 9.33 \approx 10$$

Quedando entonces los intervalos: [~,10), [10,20),[20,~]
y cada uno de los bins:

Bin1 = 0,4
Bin2 = 12,16,16,18
Bin3 = 24,26,28



Variables numéricas - Transformación

Discretización no-supervisada

Discretización de frecuencia uniforme

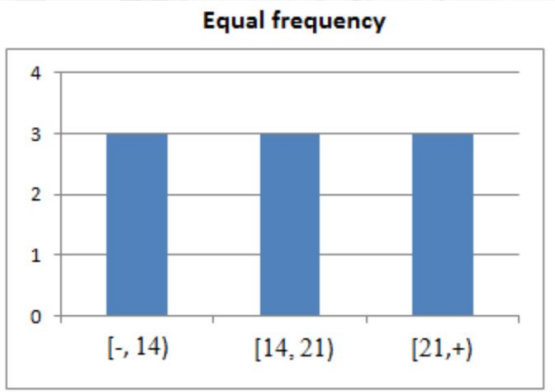
Se dividen los datos en *k* grupos en el que cada uno mantenga la misma cantidad de datos.
La mejor forma de conseguir esta discretización es a partir del histograma.

Ejemplo: Con los mismos datos anteriores {*X* = 0,4,12,16,16,18,24,26,28} y se quiere dividir en **3 intervalos**, k=3.

Quedando entonces los intervalos: [~,14), [14,21),[21,~)

Y cada uno de los Bins:

Bins1 = 0, 4, 12
Bins2 = 16,16,18
Bins3 = 24,26,28



Variables numéricas - Transformación

Discretización supervisada

- Dentro de este enfoque sí se hace uso del conocimiento de la clase o etiqueta de los datos.
- Busca la mejor partición de forma que cada grupo sea lo más **“puro”** posible.
- Entiéndase el concepto de pureza como que la mayoría de elementos en cada grupo (bin) pertenezca a la misma clase.
- Un método de este enfoque, ampliamente utilizado es el **Método de la entropía**

Variables numéricas - Transformación

Discretización supervisada

Método de la entropía

La entropía, también llamada entropía de Shannon o entropía de información es una medida de la incertidumbre de una fuente de información, o dicho de otra forma, la cantidad de información que contiene los símbolos usados.

Sea **S** un experimento aleatorio, **C** las posibles salidas de ese experimento y p_i la probabilidad de cada salida. La entropía **E(S)** se calcula como:

$$E(S) = \sum_{i=1}^{C} -p_i \log_2 p_i$$

Cuando se tiene un experimento/proceso donde todas sus posibles salidas son equiprobables, decimos que el resultado es poco predecible y por tanto la entropía será máxima. Por el contrario, si la salida es predecible su entropía será mínima.

Variables numéricas - Transformación

Discretización supervisada

Método de la entropía

Ejemplo:

Variables categóricas - Transformación

La transformación de una variable categórica se realiza por medio de un proceso conocido como codificación, encoding o continuización.

Se busca que los valores categóricos se mapean en valores numéricos.

Ejemplo: tratar esta variable como la de género como números enteros.

Femenino → 1, Masculino → 0

Los métodos más comunes son: codificación binaria y codificación usada en objetivo.

Es importante recordar que existen dos tipos de variables: **categóricas, ordinales y nominales.**

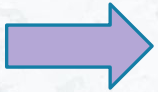
Variables categóricas - Transformación

Label Encoding - Ordinal Encoding

Utilizamos esta técnica de codificación de datos categóricos cuando la característica categórica es ordinal. En este caso, es importante conservar el orden. Por lo tanto, la codificación debe reflejar la secuencia.

En la codificación de etiquetas, cada etiqueta se convierte en un valor entero.

	Grado
0	Primaria
1	Maestría
2	Pregrado
3	Secundaria
4	Secundaria
5	PhD
6	Primaria
7	Maestría
8	Pregrado



	Grado
0	0
1	3
2	2
3	1
4	1
5	4
6	0
7	3
8	2

Variables categóricas - Transformación

	Ciudades
0	Pereira
1	Cali
2	Medellín
3	Bogotá
4	Barranquilla
5	Cartago

	Dummy
0	10000
1	01000
2	00100
3	00010
4	00001
5	00000

	Pereira	Cali	Medellín	Bogotá	Barranquilla	Cartago
0	1	0	0	0	0	0
1	0	1	0	0	0	0
2	0	0	1	0	0	0
3	0	0	0	1	0	0
4	0	0	0	0	1	0
5	0	0	0	0	0	1

Dummy Encoding - One-hot Encoding

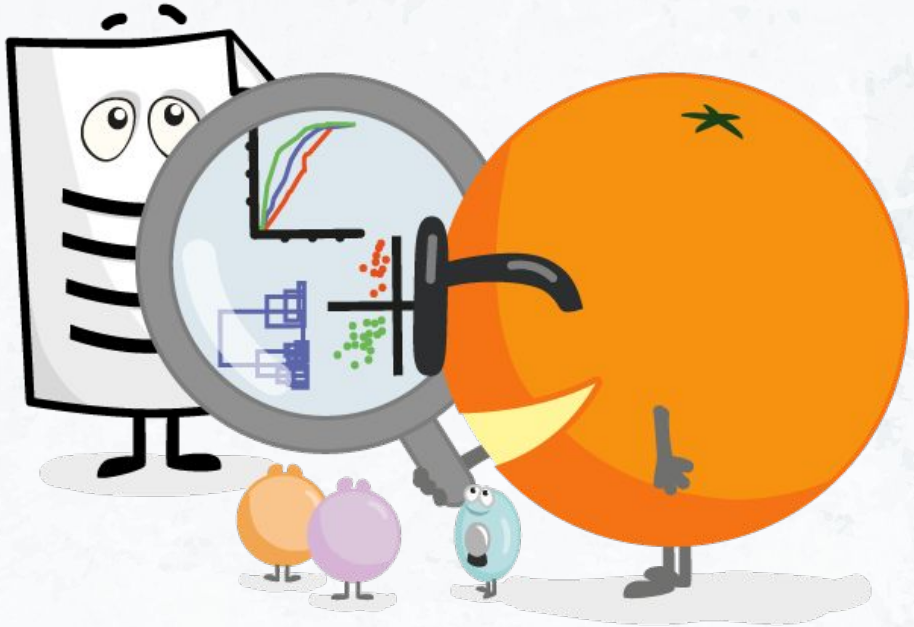
El esquema de codificación dummy es similar a la codificación one-hot.

Este método de codificación de datos categóricos transforma la variable categórica en un conjunto de variables binarias (también conocidas como variables dummies).

En el caso de la codificación one-hot, para N categorías en una variable, utiliza N variables binarias. La codificación dummy es una pequeña mejora respecto a la codificación one-hot.

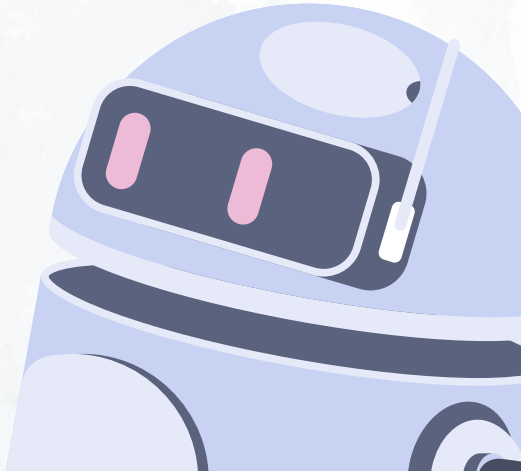
La codificación dummy utiliza N-1 características para representar N etiquetas/categorías.

Datos faltantes



Gracias !

dfcollazosh@unal.edu.co



UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD
NACIONAL
DE COLOMBIA