FUNDACIÓN YPF

Módulo III | Clase 7

Análisis exploratorio de datos: Transformación de datos

FUNDACIÓN YPF

¿Ponemos a grabar el taller?

FUNDACIÓN YPF

¿Qué vamos a ver hoy?

- Conceptos básicos de Feature Engineer
- Obtención de variables derivadas

- Encoder de variables
- Discretización de variables
- Re- escalado de variables

FUNDACIÓN YPF

Feature Engineering

FUNDACIÓN YPF

Feature Engineer

El Feature Engineer persigue como objetivo encontrar features relevantes para la pregunta que queremos contestar en los datos.

Depende de que hayamos **planteado correctamente** el problema a ser optimizado.

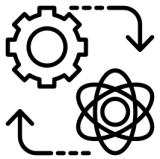
FUNDACIÓN YPF

¿Qué implica?

Selección de Features

Combinación de variables

Transformación de variables



FUNDACIÓN YPF

Obtención de variables

FUNDACIÓN YPF

Obtención de variables derivadas

Combinar variables ya presentes que puedan capturar efectos dependientes entre ambos y afectar el resultado:

- Sumando o restando variables
- Multiplicando o dividiendo variables

FUNDACIÓN YPF

Encoding de variables

FUNDACIÓN YPF

Encoding de variables

En muchos datasets tenemos **variables categóricas**. Pero pocos algoritmos puede lidiar con este tipo de variables. La mayoría de ellos espera **valores numéricos.**

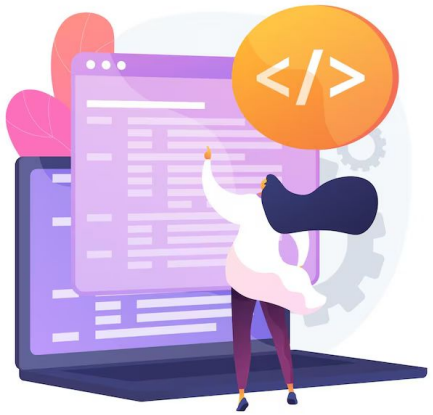
Convertimos los datos categóricos en datos numéricos.

- LabelEncoder()
- .get_dummies()
- OneHotEncoder()

FUNDACIÓN YPF

.LabelEncoder()

Codifica categoría en valores numéricos entre 0 y el número de clases menos 1.



	pais	edad	salario	compra
0	Francia	44	72000	no
1	España	27	48000	si
2	Alemania	30	54000	no
3	España	38	61000	no
4	Alemania	40	40000	si
5	Francia	35	58000	si
6	España	40	52000	no
7	Francia	48	79000	si
8	Alemania	50	83000	no
9	Francia	37	67000	si

	pais	edad	salario	compra
0	2	44	72000	no
1	1	27	48000	si
2	0	30	54000	no
3	1	38	61000	no
4	0	40	40000	si
5	2	35	58000	si
6	1	40	52000	no
7	2	48	79000	si
8	0	50	83000	no
9	2	37	67000	si

FUNDACIÓN YPF

.get_dummies()

- **Variable dummy** o “indicadora”
- Función de pandas: modifica el DataFrame, Se aplica para cada atributo por separado.

FUNDACIÓN YPF

OneHotEncoder()

- **Variable dummy** o “indicadora”
- Función de sklearn: no modifica el DataFrame

FUNDACIÓN YPF

Discretización de variables

FUNDACIÓN YPF

Discretización de variables

A veces las **variables numéricas** no aportan mucha información como variable continua. Determinar rangos y asignar categorías puede aportar más información

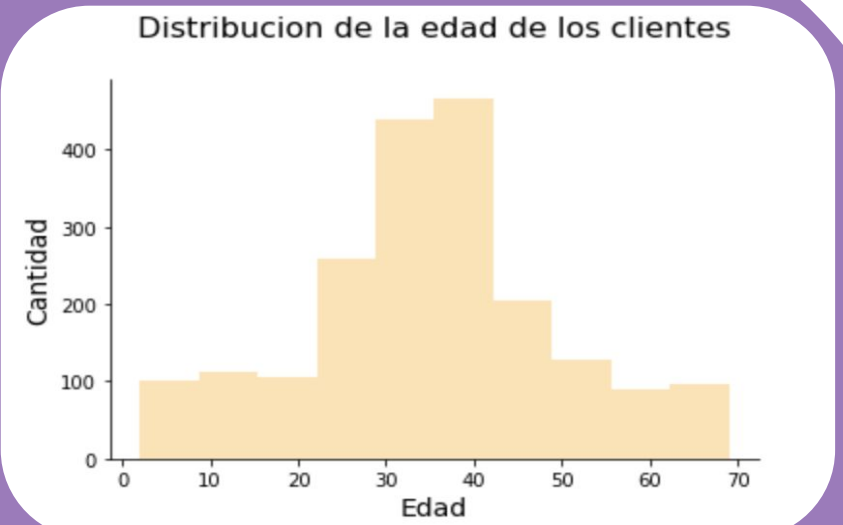
Reemplazamos los valores contenidos en un pequeño intervalo con un único valor representativo para el mismo

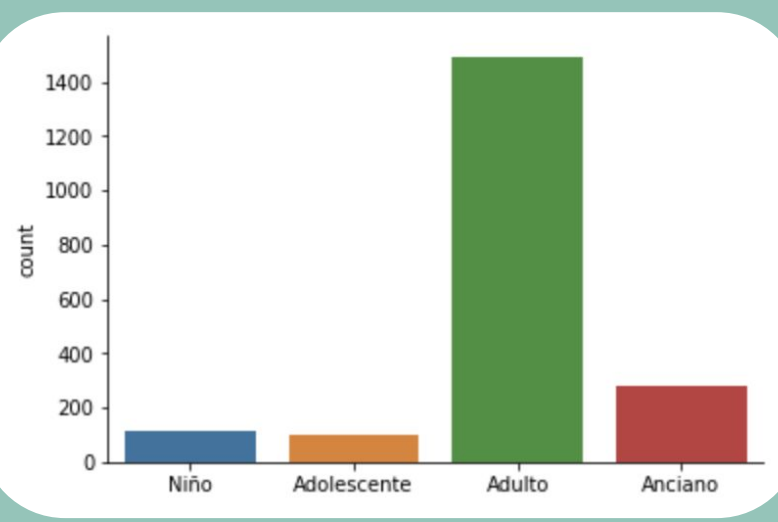
- Binning

FUNDACIÓN YPF

Discretización de variables

Binning





FUNDACIÓN YPF

Descanso

Nos vemos en 10 minutos

FUNDACIÓN YPF

Sección práctica:

Aprendemos cómo transformar datos en la Notebook 11

FUNDACIÓN YPF

Sala general:

Transformación de datos

Trabajamos con la Notebook 11

Demostraremos cómo obtener variables derivadas y como hacer encoding de las variables usando pandas y scikit-learn.

FUNDACIÓN YPF

DESAFÍO 9

Para la siguiente clase:

- comenzar a aplicar a su tema y dataset la práctica vista en clase:
- Transformar datos con Pandas

FUNDACIÓN YPF

¿Alguna consulta?

FUNDACIÓN YPF

Repasamos

FUNDACIÓN YPF

Feature Engineer

El Feature Engineer persigue como objetivo encontrar features relevantes para la pregunta que queremos contestar en los datos.

Depende de que hayamos **planteado correctamente** el problema a ser optimizado.

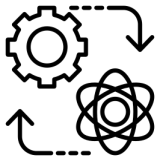
FUNDACIÓN YPF

¿Qué implica?

Selección de Features

Combinación de variables

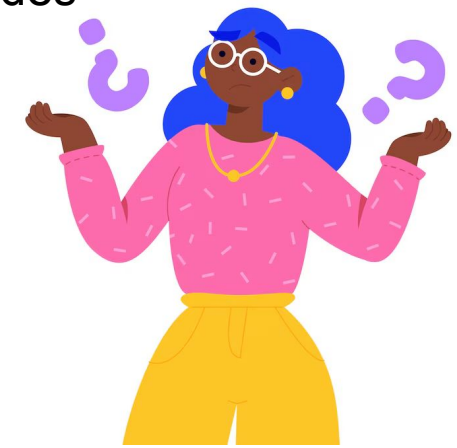
Transformación de variables



FUNDACIÓN YPF

¿Por qué transformar las variables?

Muchos algoritmos de machine learning tienen un mejor desempeño cuando las variables o features están en una **escala similar** o **distribuidas normalmente**. Cuando el rango de una variable es más pequeño, las variaciones pequeñas son importantes pero quedan cubiertas por grandes variaciones en otras variables.



FUNDACIÓN YPF

Re - escalado de variables

FUNDACIÓN YPF

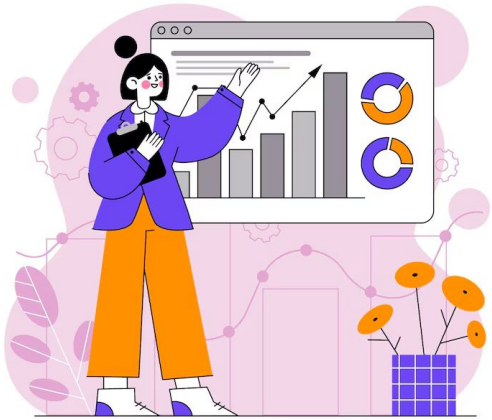
Re-escalado de variables

Se utiliza si debemos escalar los datos al rango [0,1), particularmente en algoritmos donde la distancia es importante.

Se cambia el rango de los datos

- MinMaxScaler()

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}}$$



Support vector machines (SVM) y **k-nearest neighbors (KNN)**

FUNDACIÓN YPF

¡Muchas gracias!