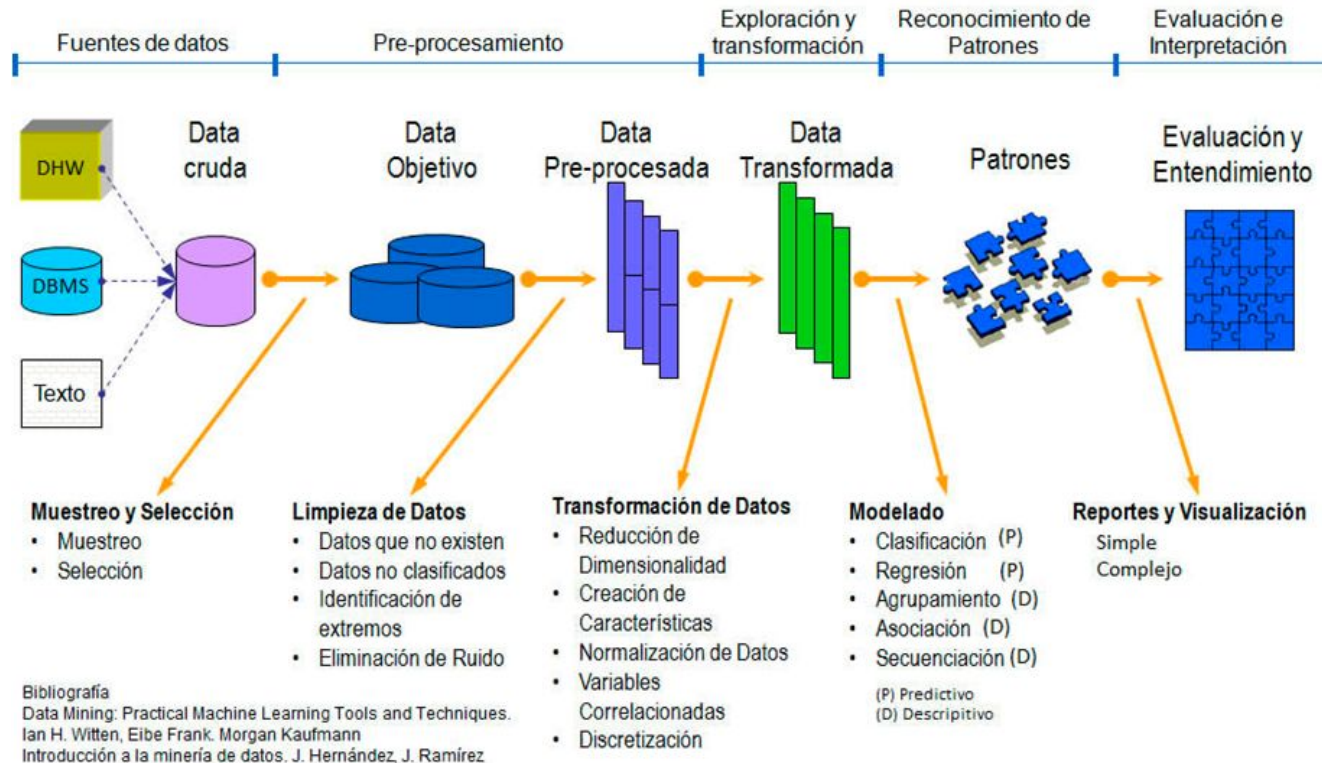


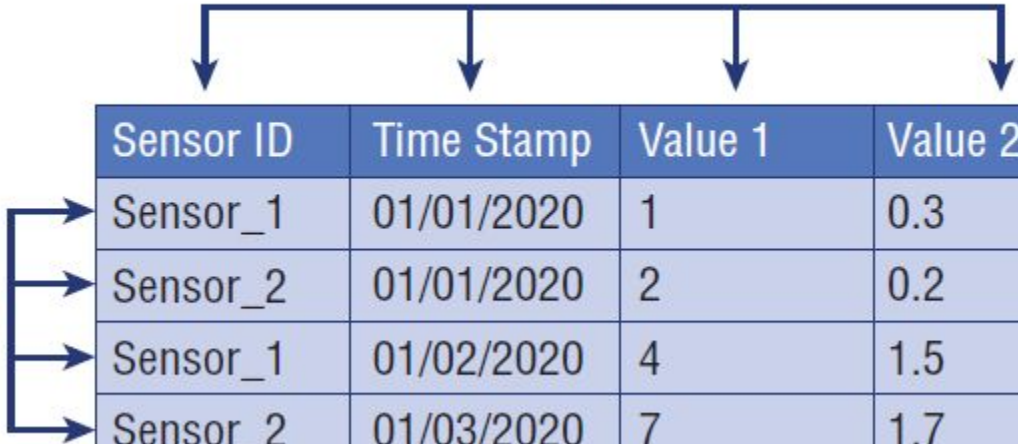
Pre-procesamiento

Angela López
Kevin Rubiano
Deivis Cárdenas

Proceso del análisis de datos



¿Qué es una característica?



The diagram illustrates a data table with four columns and four rows. The columns are labeled 'Sensor ID', 'Time Stamp', 'Value 1', and 'Value 2'. The rows represent observations from two sensors over time. Annotations include 'Columns (Features)' with arrows pointing to each column header, and 'Rows (Observations)' with arrows pointing to each row of data.

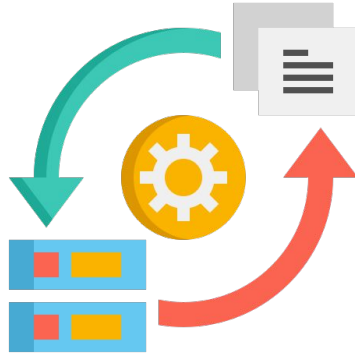
Columns (Features)			
Sensor ID	Time Stamp	Value 1	Value 2
Sensor_1	01/01/2020	1	0.3
Sensor_2	01/01/2020	2	0.2
Sensor_1	01/02/2020	4	1.5
Sensor_2	01/03/2020	7	1.7

Rows (Observations)

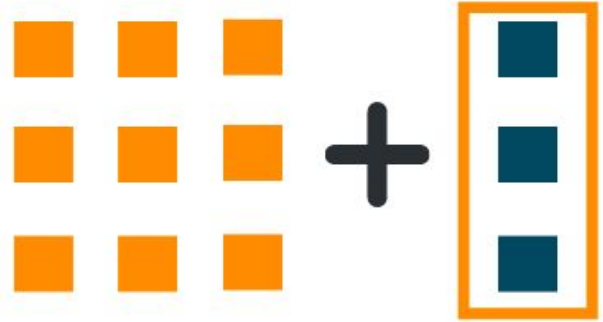
Ingeniería de características



Limpieza



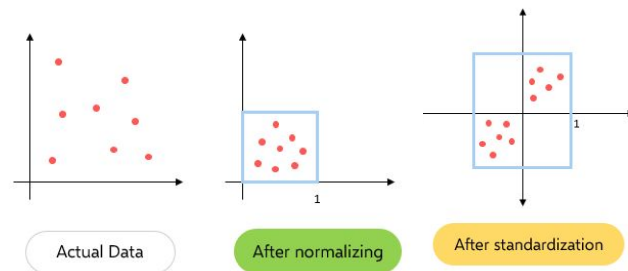
Transformación



Adición

Ingeniería de características

- Estandarización
- Imputación de datos faltantes
- Normalización
- Transformaciones no lineales
- Codificación de características categóricas
- Discretización



Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	1	0

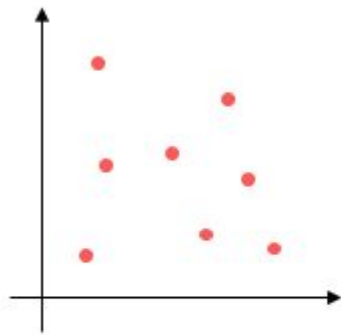
Imputación de datos faltantes

- Imputación simple: Media, mediana, por regresión, hot-deck
- Imputación múltiple

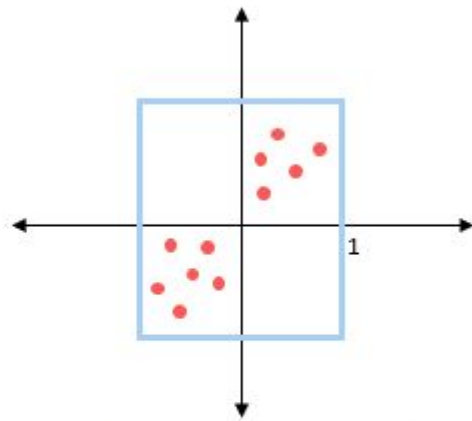


Estandarización de características

- Escalar características a un rango
- Escalar datos con valores atípicos
- Escalar datos dispersos



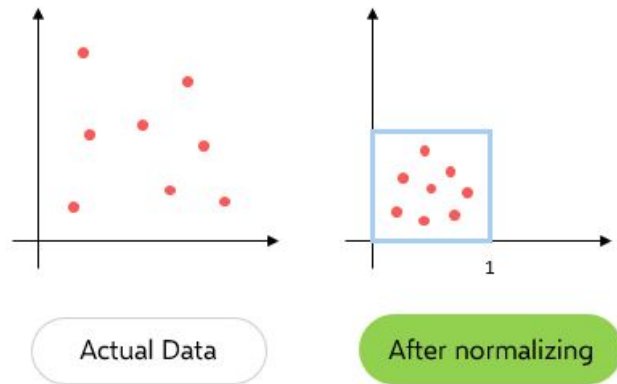
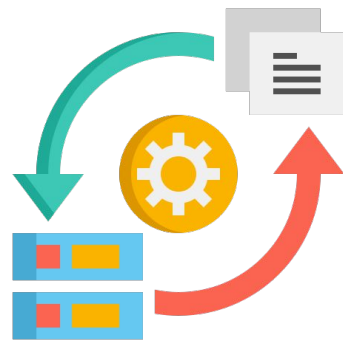
Actual Data



After standardization

Transformación de los datos

- Cuando la presentación de una variable no es adecuada para el modelo, pero la información que contiene es necesaria para el mismo.
- Mejorar las propiedades de una variable.



Normalización

- MÍN-MÁX

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{nuevo_max}_A - \text{nuevo_min}_A) + \text{nuevo_min}_A$$

- Z-SCORE

$$v' = \frac{v - \text{media}_A}{\text{des_est}_A} \quad s_A = \frac{1}{n} \sum_{i=1}^n |v_i - \text{media}_A|$$

- ESCALA DECIMAL

Transformaciones no lineales

- Buscar normalidad en la distribución de los datos.
- Estabilizar la varianza
- Minimizar la asimetría

Transformación de Box-cox

$$v' = \begin{cases} \frac{x_i - 1}{\lambda} & \lambda \neq 0 \\ \ln(x_i) & \lambda = 0 \end{cases}$$

Transformación de Yeo - Johnson

$$x_i^{(\lambda)} = \begin{cases} [(x_i + 1)^\lambda - 1]/\lambda & \text{if } \lambda \neq 0, x_i \geq 0, \\ \ln(x_i + 1) & \text{if } \lambda = 0, x_i \geq 0 \\ -[(-x_i + 1)^{2-\lambda} - 1]/(2 - \lambda) & \text{if } \lambda \neq 2, x_i < 0, \\ -\ln(-x_i + 1) & \text{if } \lambda = 2, x_i < 0 \end{cases}$$

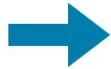
Codificación de características categóricas

Los datos categóricos representan un reto para algunas técnicas de aprendizaje de máquina ya que estas están hechas para trabajar con valores numéricos, es por esta razón que tenemos que hacer una transformación de nuestros datos categóricos a valores numéricos.

Codificación de características categóricas

La codificación de etiquetas

Color
Red
Red
Yellow
Green
Yellow



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Gender
Female
Male
Male
Female

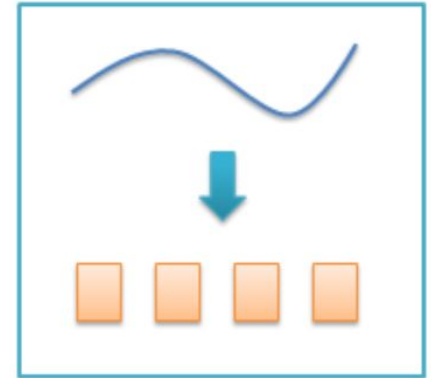


Gender
1
0
0
1

Variables Dummy

Discretización de datos

- Los datos son reemplazados por rangos o por datos con niveles conceptuales superiores.
- Reducir el número de valores de un atributo continuo, dividiendo el rango de atributos en intervalos.



Ejemplo

Cientes que
compraron y clientes
que no compraron

No	País	Edad	Salario	Compra
1	Francia	44	72000	No
2	España	27	48000	Si
3	Alemania	30	54000	No
4	España	38	61000	No
5	Alemania	40		Si
6	Francia	35	58000	Si
7	España		52000	No
8	Francia	48	79000	Si
9	Alemania	50	83000	No
10	Francia	37	67000	Si

No	País	Edad	Salario	Compra
1	Francia	44	72000	No
2	España	27	48000	Si
3	Alemania	30	54000	No
4	España	38	61000	No
5	Alemania	40	63777.77	Si
6	Francia	35	58000	Si
7	España	38.77	52000	No
8	Francia	48	79000	Si
9	Alemania	50	83000	No
10	Francia	37	67000	Si

Imputación de datos
faltantes

No	País	Edad	Salario	Compra
1	Francia	44.000000	72000.000000	0
2	España	27.000000	48000.000000	1
3	Alemania	30.000000	54000.000000	0
4	España	38.000000	61000.000000	0
5	Alemania	40.000000	63777.777778	1
6	Francia	35.000000	58000.000000	1
7	España	38.777778	52000.000000	0
8	Francia	48.000000	79000.000000	1
9	Alemania	50.000000	83000.000000	0
10	Francia	37.000000	67000.000000	1

Codificación de variables categóricas
"binarias"

Codificación de variables categóricas "dummy"

No	Edad	Salario	Compra	País_Alemania	País_España	País_Francia
1	44.000000	72000.000000	0	0	0	1
2	27.000000	48000.000000	1	0	1	0
3	30.000000	54000.000000	0	1	0	0
4	38.000000	61000.000000	0	0	1	0
5	40.000000	63777.777778	1	1	0	0
6	35.000000	58000.000000	1	0	0	1
7	38.777778	52000.000000	0	0	1	0
8	48.000000	79000.000000	1	0	0	1
9	50.000000	83000.000000	0	1	0	0
10	37.000000	67000.000000	1	0	0	1

No	Edad	Salario	Compra	País_Alemania	País_España	País_Francia
1	0.739130	0.685714	0	0	0	1
2	0.000000	0.000000	1	0	1	0
3	0.130435	0.171429	0	1	0	0
4	0.478261	0.371429	0	0	1	0
5	0.565217	0.450794	1	1	0	0
6	0.347826	0.285714	1	0	0	1
7	0.512077	0.114286	0	0	1	0
8	0.913043	0.885714	1	0	0	1
9	1.000000	1.000000	0	1	0	0
10	0.434783	0.542857	1	0	0	1

Normalización de
características

Paquetes

- Scikit-learn
- Feature-engine
- Pandas