

Data Mining



Feature Engineering



Temas

- ❑ Concepto de Feature Engineering.
- ❑ Métodos de construcción de variables por:
 - ❑ Discretización, normalización y binning.
- ❑ Evaluación de las transformaciones.

Feature Engineering

- ❑ Dentro del Proceso de Descubrimiento de Conocimiento se corresponde con la etapa de **Transformación de datos**.
- ❑ La transformación de datos engloba, en realidad, cualquier proceso que modifique la forma de los datos.
- ❑ Prácticamente todos los procesos de preparación de datos entrañan algún tipo de transformación.

❑ ***Feature Engineering*** es la tarea de **mejorar el rendimiento** del modelado en un conjunto de datos mediante la **transformación** de su ***feature space***.

Normalización

La normalización consiste en **escalar los features** (numéricos) de manera que puedan ser mapeados a un rango más pequeño.

Por ejemplo: 0 a 1 ó -1 a 1.

La normalización es particularmente utilizada en:

- ❑ Tareas de mining donde las unidades de medidas dificultan la comparación de features.
- ❑ Medidas de Distancias. Vecinos más cercanos, Clustering, etc.

Ayuda a evitar que atributos con mayores magnitudes tengan mayor peso que los rangos pequeños.

Los métodos más utilizados para normalizar son:

- ❑ Min-Max
- ❑ Z-Score
- ❑ Decimal Scaling

Normalización Min-Max

La **Normalización Min-Max** funciona al ver cuánto más grande es el valor actual del valor mínimo **$\min(X)$** y escala esta diferencia por el rango.


$$X_{mm}^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Los valores de normalización min-max van de 0 a 1.

Ejemplo: Dataset Iris y variable Sepal.Length

Sepal.Length

Min.	4.300
1st Qu.	5.100
Median	5.800
Mean	5.843
3rd Qu.	6.400
Max.	7.900


$$X_{mm} = \frac{X - 4.3}{7.9 - 4.3}$$

Para los valores extremos es 0 y 1

Normalización Z-Score

Los valores para un **atributo X**, se normalizan en base a la **media** y **desviación estándar** de **X**.

$$Z-score = \frac{X - mean(X)}{sd(X)}$$

Este método de normalización es útil cuando el verdadero mínimo y máximo del atributo X son desconocidos, o cuando hay valores atípicos que dominan la normalización min-max.

Sepal.Length

Min.	4.300
Median	5.800
Mean	5.843
Max.	7.900
SD	0.828

Para una Iris con el largo del sépalo más corto: $Z-score = \frac{4.3 - 5.843}{0.828} = -1,863$

Para una Iris con el largo del sépalo más largo: $Z-score = \frac{7.9 - 5.843}{0.828} = 2,484$

Normalización Decimal Scaling

Decimal Scaling asegura que cada valor normalizado se encuentra entre - 1 y 1.

$$X_{decimal} = \frac{X}{10^d}$$

donde **d** representa el número de dígitos en los valores de la variable con el valor absoluto más grande.

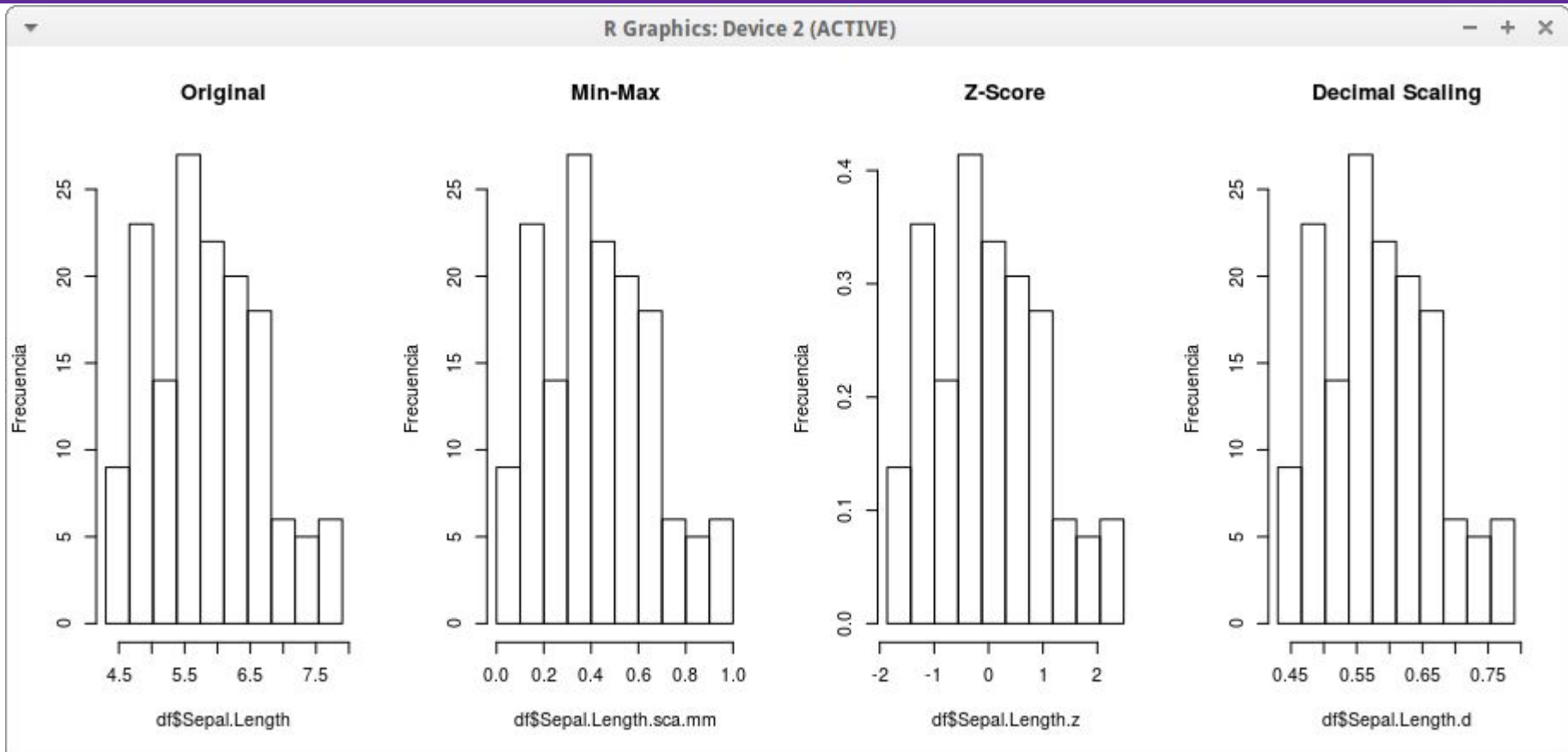
Sepal.Length

Min.	4.300
Median	5.800
Mean	5.843
Max.	7.900
SD	0.828

$$X_{decimal} = \frac{4.3}{10^1}$$

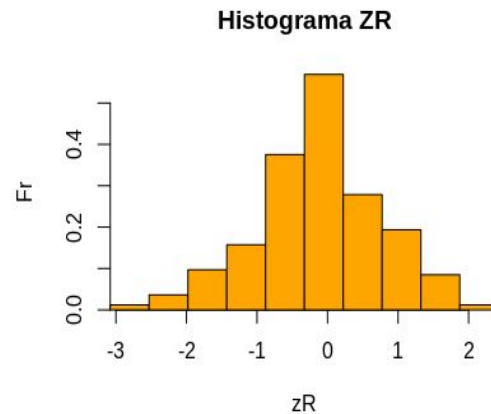
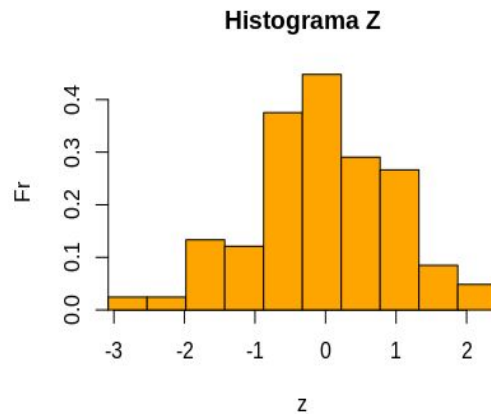
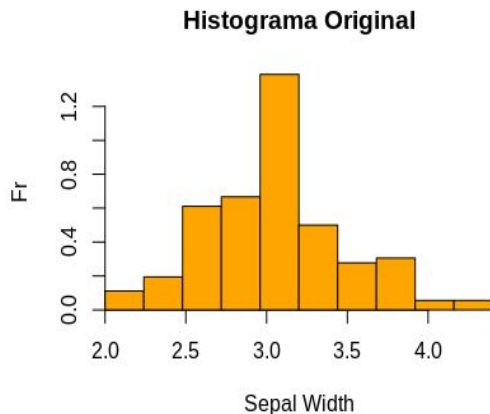
$$X_{decimal} = \frac{7.9}{10^1}$$

Comparación de métodos



Escalados Robustos

- ❑ Si nuestro dataset contienen muchos valores atípicos, es probable que un escalado utilizando la media y la varianza de los datos no funcione muy bien.
- ❑ En estos casos, puede usar un método *robusto* como reemplazo.
 - ❑ Usan estimaciones más sólidas para el **centro** y el **rango** de sus datos.
 - ❑ Por ejemplo: Mediana (o algún percentil) e IQR



Transformaciones para lograr Normalidad

$$Sesgo = \frac{3*(media-mediana)}{desví o}$$

- ❑ Si la media es mayor que la mediana entonces hay sesgo a derecha (Sesgo+)
- ❑ Si la media es menor que la mediana entonces hay sesgo a izquierda (Sesgo-)

Los histogramas de la diapo anterior, existía un leve sesgo positivo:

Podemos reducir este sesgo a partir de transformaciones:

1. Raíz cuadrada
2. Logaritmos
3. Inversa de la Raíz Cuadrada

```
> print(sesgo.tr.sq)
[1] 0.052761
> print(sesgo.tr.ln)
[1] -0.05237737
```

```
> print(sesgo.ori)
[1] 0.1569923
> print(sesgo.mm)
[1] 0.1569923
> print(sesgo.z)
[1] 0.1569923
> print(sesgo.d)
[1] 0.1569923
```

Discretización

- ❑ Es una técnica que permite dividir el rango de una variable continua en intervalos.
- ❑ Vamos de valores continuos a un número reducido de etiquetas.
- ❑ Esto conduce a una representación concisa y fácil de utiliza.

Discretización: Características

La discretización puede ser caracterizada según cómo se realiza:

- ❑ Si utiliza la clase como información
 - ❑ Si la utiliza será **supervisada**
 - ❑ Si no la utiliza será **no supervisada**

- ❑ Según la orientación en la que realice las recursivas particiones:
 - ❑ **Top-Down**: Parte de algunos pocos puntos de *splitting* y trata de separar todo el rango
 - ❑ **Bottom-Up**: Considera a todos los puntos como posibles separadores del rango

Si se realiza una discretización para cada sub-rango de manera recursiva se obtiene una Jerarquía.

Discretización: Binning

- ❑ La técnica es similar a la que utilizamos para manejo de ruido: suavizados.
- ❑ Es **Top-Down**.
- ❑ Se basa en un número específico de **bins**.
- ❑ Los criterios de agrupamiento pueden ser por:
 - ❑ Igual-Frecuencia: La misma cantidad de observaciones en un bin.
 - ❑ Igual-Ancho: Definimos rangos o intervalos de clases para cada bin.

A su vez para cada uno de los agrupamientos podemos hacer:

- ❑ Reemplazo por **media**
- ❑ Reemplazo por **mediana**
- ❑ O una etiqueta (valor entero)

No se utiliza la información de la clase, por lo tanto es **no supervisado**.

Binning: Ejemplos

Supongamos que vamos a discretizar **X** en 3 categorías.

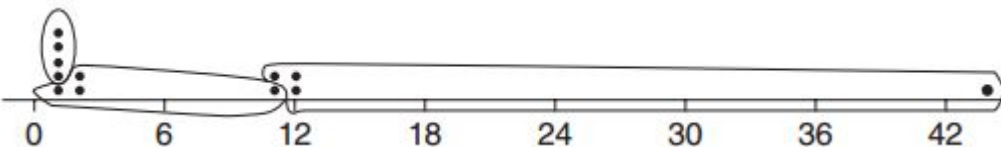
$X = \{1, 1, 1, 1, 1, 2, 2, 11, 11, 12, 12, 44\}$

Igual ancho



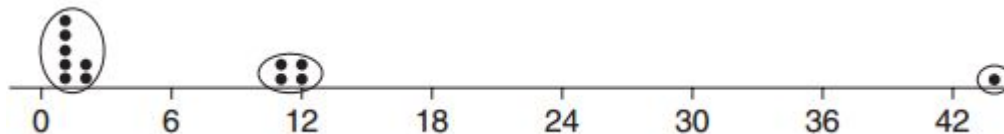
Bajo: $0 \leq X < 15$
Medio: $15 \leq X < 30$
Alto: $30 \leq X < 45$

Igual frecuencia



$n = 12$
 $\text{bins} = 3$
 $n/\text{bins} = 4$

K-means



Identifica lo que parece
ser la partición
intuitivamente correcta

Discretización: Otros no supervisados

- ❑ **Rank:** El ranking de un número **es su tamaño relativo a otros valores** de una variable numérica. Primero, ordenamos la lista de valores, luego asignamos la posición de un valor como su rango.
- ❑ **Los mismos valores reciben el mismo rango** pero la presencia de valores duplicados afecta a las filas de valores posteriores (por ejemplo, 1,2,3,3,4).
- ❑ Rango es un sólido método de binning con un inconveniente importante, los valores pueden tener rangos diferentes en diferentes listas.

Discretización: Otros no supervisados

- ❑ **Quantiles** (median, quartiles, percentiles, ...): **Quantiles** también son métodos binning muy útiles pero como Rank, un valor puede tener cuantil diferente si la lista de valores cambia.
- ❑ **Math functions**: Por ejemplo, $\text{FLOOR}(\text{LOG}(X))$ es un método binning efectivo para las variables numéricas con distribución altamente sesgada (por ejemplo, ingreso).

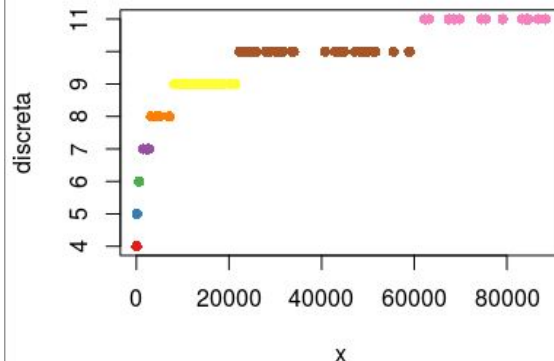
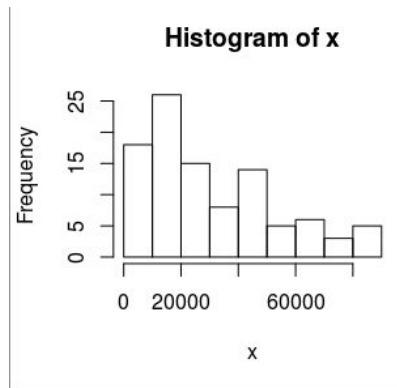
```
x = sample(10:100000, size = 100, replace = T, prob = seq(1.0, 0.0001, -0.00001))
```

```
summary(x)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
69	12958	23680	30513	44262	88312

```
unique(floor(log(x)))
```

```
[1] 11 10 8 9 5 7 6 4
```



Discretización basada en Entropía

- ❑ Es supervisada y Separación Top-Down
- ❑ Explora la distribución de información en la clase para el cálculo y determinación del **split-point**
- ❑ Para un dataset $D \rightarrow \{A_1, \dots, A_N\}$ el método para discretizar A es:
 1. Cada **Valor de A** se considera como un posible **split-point** para hacer una discretización **binaria**.
 2. Calculo la **Entropía** para la Clase

$$E(S) = \sum -p_i \ln p_i$$

3. Calcula la **Entropía** para la Clase y el **split-point** a evaluar

$$E(S, A) = \sum \frac{|S_v|}{|S|} E(S_v)$$

4. Calculo **Information Gain** para esa partición, como:

$$InformationGain = E(S) - E(S, A)$$

Discretización basada en Entropía: Ejemplo

Discretizar la variable de temperatura usando el algoritmo basado en entropía.

O-Ring Failure	
Y	N
7	17

Paso 1: Calculamos Entropía para la variable objetivo.

$$E(\text{Failure}) = E(7, 17) = -0.29 * \log_2(0.29) - 0.71 * \log_2(0.71) = 0.871$$

Paso 2: Calculamos Entropía para la variable objetivo dado un bin.

$$\begin{aligned} E(\text{Failure}, \text{Temperature}) &= P(\leq 60) * E(3, 0) + P(> 60) * E(4, 17) = \\ &= \frac{3}{24} * 0 + \frac{21}{24} * 0.7 = 0.615 \end{aligned}$$

Paso 3: Calculamos Ganancia de Información} (GI) dado un bin.

$$GI = E(S) - E(S, A)$$

$$GI(\text{Failure}, \text{Temperature}) = 0.256$$

		O-Ring Failure	
		Y	N
Temperature	≤ 60	3	0
	> 60	4	17

Variables Flags

- ❑ Algunos métodos analíticos, como la regresión, requieren que los **predictores sean numéricos**.
- ❑ Cuando tenemos descriptores categóricos, podemos recodificar la variable categórica en una o más **variables Dummy o Flags**.

Variables con dos categorías

```
If sex = female then sex_flag = 0;  
if sex = male then sex_flag = 1.
```

Variables con N categorías

```
north_flag:   If region = north then north_flag = 1; otherwise north_flag = 0.  
east_flag:    If region = east then east_flag = 1; otherwise east_flag = 0.  
south_flag:   If region = south then south_flag = 1; otherwise south_flag = 0.
```

Bibliografía

- ❑ Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., & Turaga, D. (2017). Learning feature engineering for classification. *IJCAI International Joint Conference on Artificial Intelligence*, 2529–2535. [[pdf](#)]
- ❑ Jiawei Han, Micheline Kamber, Jian Pei. 2012. Tercera edición. Data Mining: Concepts and Techniques. Cap. 2 y Cap. 3
- ❑ Daniel T. Larose. 2014. Segunda edición. Discovering Knowledge in Data: An Introduction to Data Mining.
- ❑ Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1), 3–14. [[pdf](#)]
- ❑ Should I normalize/standardize/rescale the data? [[link](#)]