Machine Learning

Unidad # 1 - Fundamentos de Machine Learning y Preparación de Datos

CC57 – 2019-1

Profesor
Andrés Melgar

Competencias a adquirir en la sesión

- Al finalizar la sesión el alumno discutirá las diferentes tareas que se se deben realizar en las fases de **selección**, **pre-procesamiento** y **transformación** del proceso de descubrimiento de conocimiento en base de datos.

- Al finalizar la sesión el alumno ejecutará tareas de **transformación de datos** considerando la **reducción de datos**.

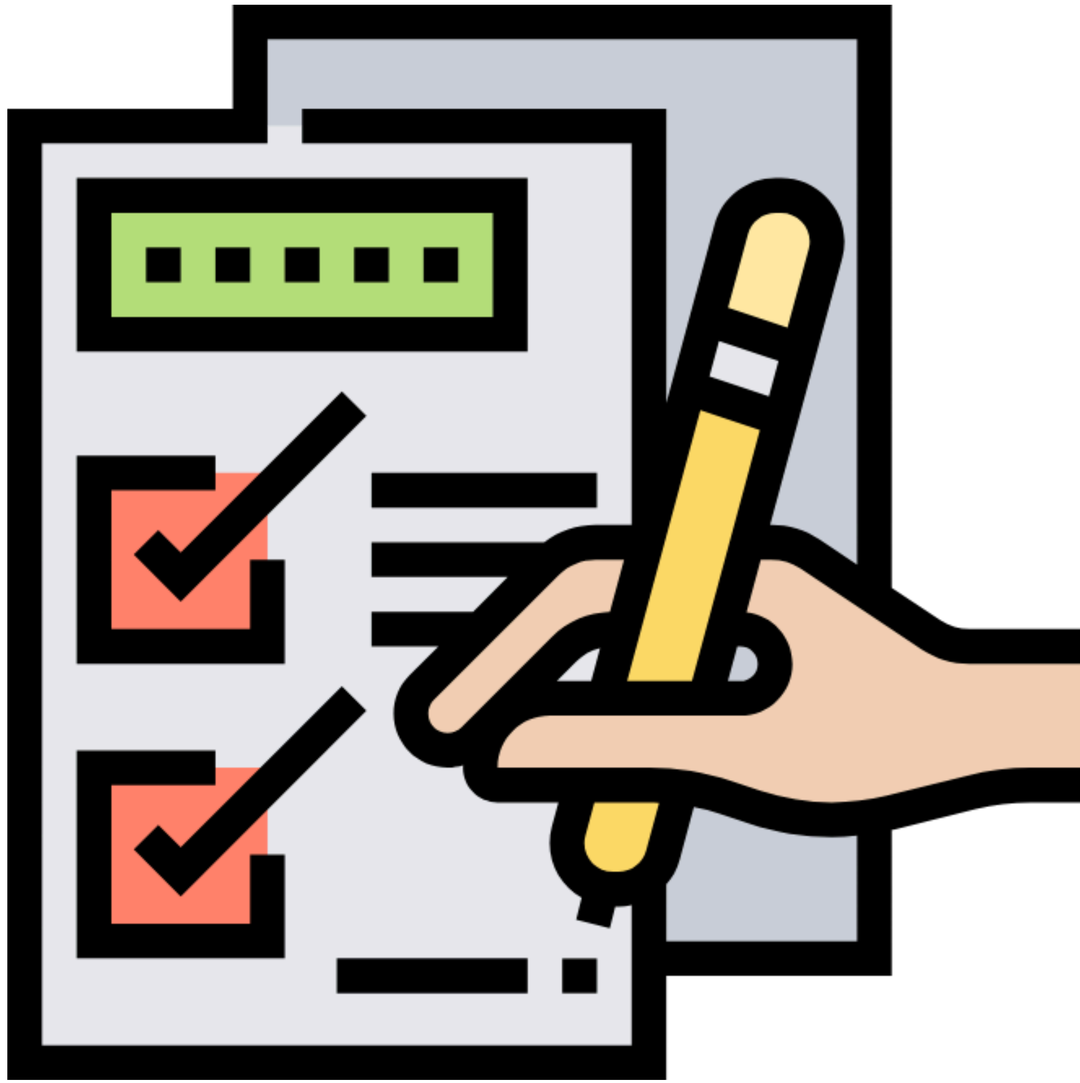


Imagen obtenida a través de la URL https://www.flaticon.com/free-icon/checklist_1467445

Revisión de la sesión anterior

- ¿Por qué es importante **gestionar los valores ausentes**?
- ¿Por qué es importante **suavizar el ruido**?
- ¿Por qué es importante eliminar los **registros duplicados**?

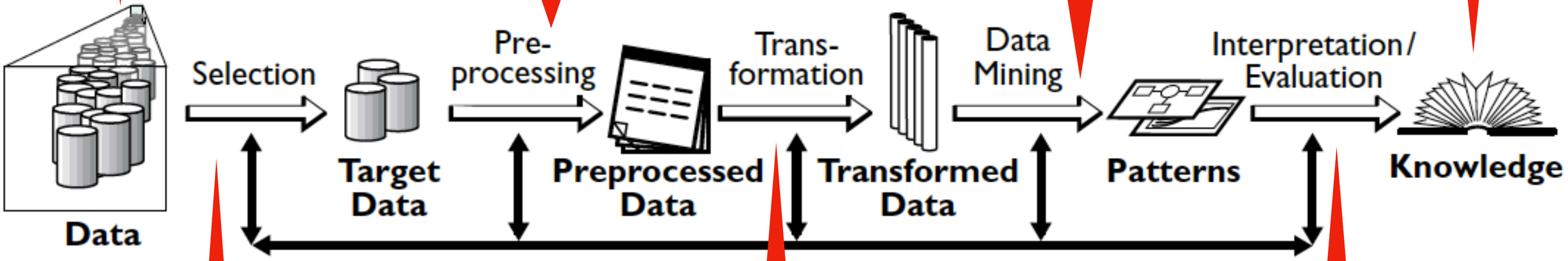
Descubierta de Conocimiento en Base de Datos

Conocimiento del dominio de la aplicación.

Limpieza de datos y preprocesamiento. Remover ruido, anomalías, manejar campos nulos, ventana de tiempo.

Selección del mejor algoritmo de aprendizaje. Depende del paradigma de aprendizaje y de la tarea a ejecutar.

Incorporar el conocimiento en un sistema.



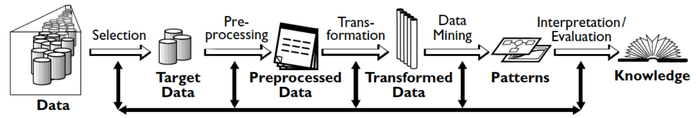
Seleccionar la base de datos o subconjunto de variables de los datos o muestra de datos.

Reducción de datos y de variables (reducción de la dimensionalidad).

Se interpreta el modelo extraído. Se evalúa el modelo para ver si tiene un rendimiento adecuado.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

Reducción de datos

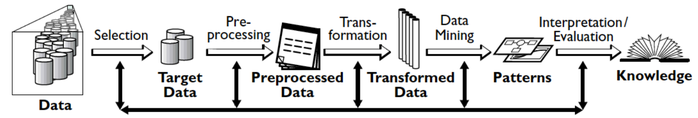


Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- El análisis de datos complejos y la minería en grandes cantidades de datos pueden llevar mucho tiempo, haciendo que ese análisis sea impráctico o inviable.
- Las técnicas de reducción de datos permiten obtener una **representación reducida** del conjunto de datos:
  - Mucha más **pequeño** en volumen.
  - Pero que mantiene la **integridad** de los datos originales.
  - Produce los **mismos resultados** analíticos (o casi los mismos).
- Entre las estrategias para reducción de datos podemos mencionar:
  - **Reducción de la dimensionalidad:** busca reducir el número de variables aleatorias o atributos bajo consideración (por ejemplo, *transformadas wavelet, PCA, selección de subconjuntos de atributos*).
  - **Reducción de volumen:** busca reducir el volumen del conjunto de datos (por ejemplo, *métodos paramétricos, histogramas, agrupación, muestreo, agregación de cubos de datos*).

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad transformadas wavelet

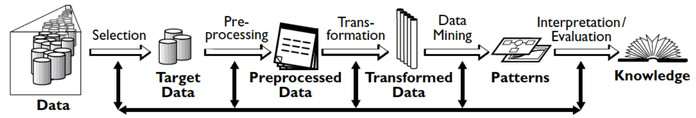


Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- La transformada discreta de wavelet (DWT) es una técnica de procesamiento de señales lineales que dado un vector X, lo **transforma** en un vector numérico X' diferente de coeficientes de wavelets.
- Los dos vectores (X, X') tienen la misma longitud.
- **¿Cómo se usa la DWT en la reducción de la dimensionalidad?**
  - Los datos de la transformada de wavelet puede ser **truncados**.
  - Una aproximación reducida de los datos se puede conseguir usando una pequeña fracción de los **coeficientes** de wavelet **más fuertes**.
  - Por ejemplo, se pueden conservar los coeficientes de wavelet mayores que algún **umbral** especificado por el usuario. Los otros coeficientes se les asigna 0.

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

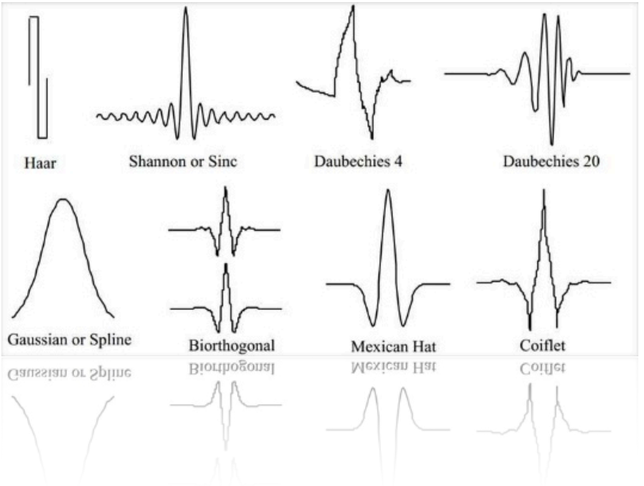
Reducción de datos
Reducción de la dimensionalidad transformadas wavelet



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

Imagen obtenida de la URL https://www.handerella.com/discrete-wavelet-transforms-theory-and-applications.html

- El procedimiento general para aplicar una DWT utiliza un **algoritmo piramidal jerárquico** que divide a la mitad los datos en cada iteración.
- El método es como sigue:

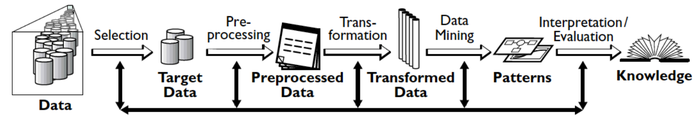


1. La **longitud,** *L*, del vector de datos de entrada debe ser una **potencia entera de 2**. Esta condición se puede cumplir rellenando el vector de datos con ceros según sea necesario.
2. Cada transformación implica **aplicar dos funciones**. La primera aplica un suavizamiento de datos (como una suma o promedio ponderado). La segunda realiza una diferencia ponderada, que actúa para resaltar las características detalladas de los datos.
3. Las dos funciones se aplican a **pares de puntos** de datos en *X*. Esto da como resultado dos conjuntos de datos de longitud *L=2*. Estos representan: i) una versión suavizada o de baja frecuencia de los datos y ii) un contenido de alta frecuencia.
4. Las dos funciones se **aplican recursivamente** a los conjuntos de datos obtenidos en el bucle anterior, hasta que los conjuntos de datos resultantes obtenidos sean de longitud *2*.
5. Los valores seleccionados de los conjuntos de datos obtenidos en las iteraciones anteriores se designan los **coeficientes de wavelet** de los datos transformados.

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos

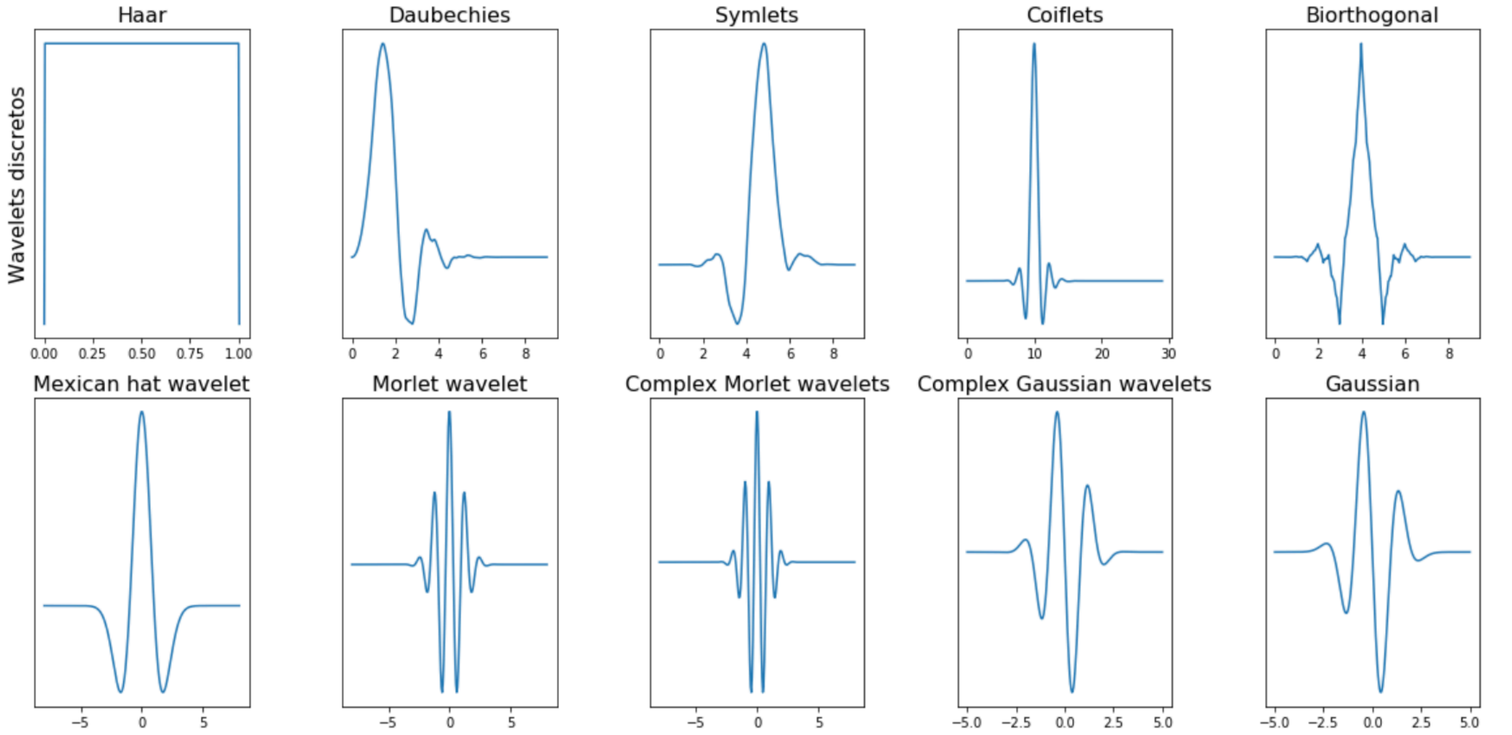
Reducción de la dimensionalidad transformadas wavelet



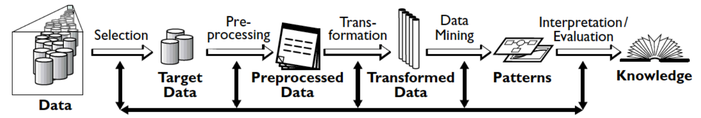
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

Imagen obtenida de la URL https://www.handerella.com/discrete-wavelet-transforms-theory-and-applications.html

```
import pandas
import pywt
import matplotlib.pyplot as plot

wavelets_discretas = ['haar', 'db5', 'sym5', 'coif5', 'bior2.4']
wavelets_continuas = ['mexh', 'morl', 'cmor1.5-1.0', 'cgau5', 'gaus5']

lista_de_listas_de_wavelets = [wavelets_discretas, wavelets_continua
lista_de_objetos = [pywt.Wavelet, pywt.ContinuousWavelet]

fig, axarr = plot.subplots(nrows=2, ncols=5, figsize=(16,8))
for ii, lista_de_wavelets in enumerate(lista_de_listas_de_wavelets):
    objeto_wavelet = lista_de_objetos[ii]
    numero_fila = ii
    for numero_columna, nombre_wavelet in enumerate(lista_de_wave
        wavelet = objeto_wavelet(nombre_wavelet)
        familia_wavelet = wavelet.family_name
        if ii == 0:
            _ = wavelet.wavefun()
            funcion_wavelet = _[0]
            valores_de_x = _[-1]
        else:
            funcion_wavelet, valores_de_x = wavelet.wavefun()
        if numero_columna == 0 and ii == 0:
            axarr[numero_fila, numero_columna].set_ylabel("Wavelets discretos", fontsize=16)
        if numero_columna == 0 and ii == 1:
            axarr[numero_fila, numero_columna].set_ylabel("Wavelets continuos", fontsize=16)
        axarr[numero_fila, numero_columna].set_title("{}".format(familia_wavelet), fontsize=16)
        axarr[numero_fila, numero_columna].plot(valores_de_x, funcion_wavelet)
        axarr[numero_fila, numero_columna].set_yticks([])
        axarr[numero_fila, numero_columna].set_yticklabels([])

plot.tight_layout()
plot.show()
```



Reducción de datos
Reducción de la dimensionalidad
PCA



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

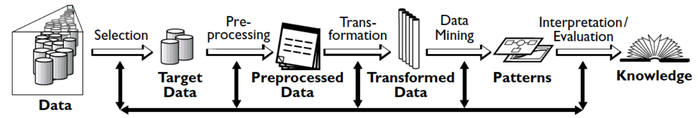
- Supongamos que los datos a reducir consisten en tuplas o vectores de datos descritos por *n* atributos o dimensiones.
- El análisis de componentes principales (PCA) busca **k vectores ortogonales** n-dimensionales que pueden representar mejor los datos, donde **k ≤ n**.
  - *Recuerde que dos vectores son ortogonales si su producto escalar es igual a cero*.
- Los datos originales se proyectan así en un **espacio** mucho **más pequeño**, lo que resulta en una reducción de la dimensionalidad.
- PCA "combina" la esencia de los atributos al crear un conjunto alternativo de variables más pequeño. Los datos iniciales se pueden **proyectar en este conjunto** más pequeño.
- El PCA a menudo revela relaciones que no se sospechaban previamente y, por lo tanto, permite interpretaciones que normalmente no darían resultado.

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos

Reducción de la dimensionalidad

PCA



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

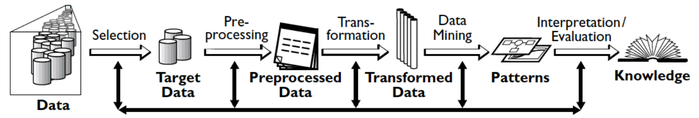
- Se tiene una regadera en 3 dimensiones, pero solo se posee fotos en 2 dimensiones. **¿Qué imagen (en 2 dimensiones) representa mejor a la regadera (3 dimensiones)?**

Hackeling, Gavin. ***Mastering Machine Learning with scikit-learn***, Packt Publishing Ltd, 2014.

Reducción de datos

Reducción de la dimensionalidad

PCA: procedimiento



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

1. Los datos de entrada se normalizan, de modo que cada atributo se encuentra dentro del mismo rango. Este paso ayuda a garantizar que los atributos con dominios grandes no dominen los atributos con dominios más pequeños.

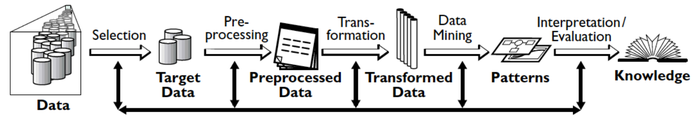
2. PCA calcula *k* vectores ortonormales que proporcionan una base para los datos de entrada normalizados. Estos son vectores unitarios en donde cada uno apunta en una dirección perpendicular a los otros. Estos vectores se conocen como los componentes principales. Los datos de entrada son una combinación lineal de los componentes principales.

3. Los componentes principales se clasifican en orden decreciente de "significación" o fuerza. Los componentes principales sirven esencialmente como un nuevo conjunto de ejes para los datos, que proporcionan información importante sobre la varianza. Es decir, los ejes ordenados son tales que el primer eje muestra la mayor varianza entre los datos, el segundo eje muestra la siguiente varianza más alta, y así sucesivamente.

4. Debido a que los componentes se clasifican en orden decreciente de "importancia", el tamaño de los datos puede reducirse eliminando los componentes más débiles, es decir, aquellos con baja varianza. Usando los componentes principales más fuertes, debería ser posible reconstruir una buena aproximación de los datos originales.

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
PCA



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

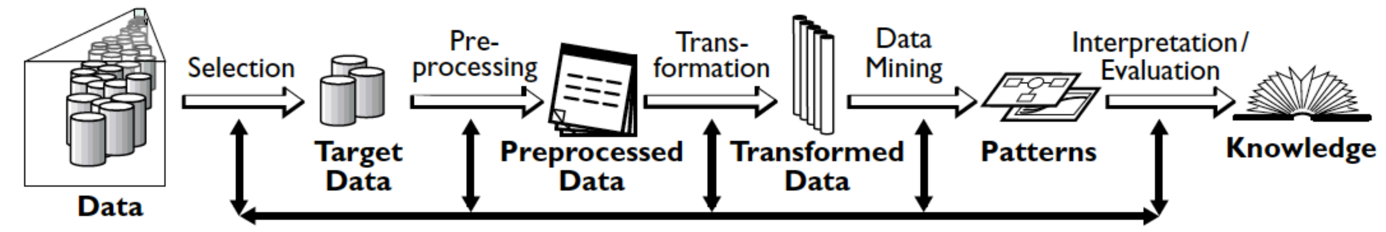
```
import pandas

archivo="iris.data"
columnas=['longitud-sépalo', 'ancho-sépalo', 'longitud-pétalo', 'ancho-pétalo', 'clase']
conjunto_de_datos = pandas.read_csv(archivo, names=columnas)

print(conjunto_de_datos.head(10))
print(conjunto_de_datos.tail(10))
```

```
   longitud-sépalo  ancho-sépalo  longitud-pétalo  ancho-pétalo        clase
0              5.1           3.5              1.4           0.2  Iris-setosa
1              4.9           3.0              1.4           0.2  Iris-setosa
2              4.7           3.2              1.3           0.2  Iris-setosa
3              4.6           3.1              1.5           0.2  Iris-setosa
4              5.0           3.6              1.4           0.2  Iris-setosa
5              5.4           3.9              1.7           0.4  Iris-setosa
6              4.6           3.4              1.4           0.3  Iris-setosa
7              5.0           3.4              1.5           0.2  Iris-setosa
8              4.4           2.9              1.4           0.2  Iris-setosa
9              4.9           3.1              1.5           0.1  Iris-setosa
```

```
     longitud-sépalo  ancho-sépalo  longitud-pétalo  ancho-pétalo  \
140              6.7           3.1              5.6           2.4
141              6.9           3.1              5.1           2.3
142              5.8           2.7              5.1           1.9
143              6.8           3.2              5.9           2.3
144              6.7           3.3              5.7           2.5
145              6.7           3.0              5.2           2.3
146              6.3           2.5              5.0           1.9
147              6.5           3.0              5.2           2.0
148              6.2           3.4              5.4           2.3
149              5.9           3.0              5.1           1.8

            clase
140  Iris-virginica
141  Iris-virginica
142  Iris-virginica
143  Iris-virginica
144  Iris-virginica
145  Iris-virginica
146  Iris-virginica
147  Iris-virginica
148  Iris-virginica
149  Iris-virginica
```

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
PCA



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *The KDD process for extracting useful knowledge from volumes of data.* Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

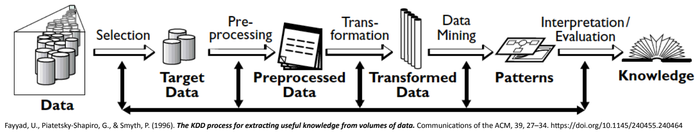
```
import pandas

#separamos los atributos del atributo meta
x = conjunto_de_datos.iloc[:,0:4].values
y = conjunto_de_datos.iloc[:,4].values

print(x)
print(y)
```

```
[['Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-versicolor' 'Iris-versicolor' 'Iris-versicolor' 'Iris-versicolor'
 'Iris-versicolor' 'Iris-versicolor' 'Iris-versicolor' 'Iris-versicolor'
 'Iris-versicolor' 'Iris-versicolor' 'Iris-versicolor' 'Iris-versicolor'
```

```
[[5.1 3.5 1.4 0.2]
 [4.9 3.  1.4 0.2]
 [4.7 3.2 1.3 0.2]
 [4.6 3.1 1.5 0.2]
 [5.  3.6 1.4 0.2]
 [5.4 3.9 1.7 0.4]
 [4.6 3.4 1.4 0.3]
 [5.  3.4 1.5 0.2]
 [4.4 2.9 1.4 0.2]
 [4.9 3.1 1.5 0.1]
 [5.4 3.7 1.5 0.2]
 [4.8 3.4 1.6 0.2]
 [4.8 3.  1.4 0.1]
 [4.3 3.  1.1 0.1]
```

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
PCA



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

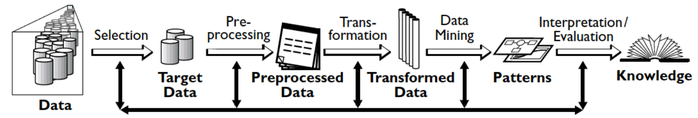
```
import pandas
from sklearn.preprocessing import StandardScaler

#estandarizamos las características
print(x)
x = StandardScaler().fit_transform(x)
print(x)
```

```
[[5.1 3.5 1.4 0.2]
 [4.9 3.  1.4 0.2]
 [4.7 3.2 1.3 0.2]
 [4.6 3.1 1.5 0.2]
 [5.  3.6 1.4 0.2]
 [5.4 3.9 1.7 0.4]
 [4.6 3.4 1.4 0.3]
 [5.  3.4 1.5 0.2]
 [4.4 2.9 1.4 0.2]
 [4.9 3.1 1.5 0.1]
 [5.4 3.7 1.5 0.2]
 [4.8 3.4 1.6 0.2]
 [4.8 3.  1.4 0.1]
 [4.3 3.  1.1 0.1]
```

```
[[-9.00681170e-01  1.03205722e+00 -1.34127240e+00 -1.31297673e+00]
 [-1.14301691e+00 -1.24957601e-01 -1.34127240e+00 -1.31297673e+00]
 [-1.38535265e+00  3.37848329e-01 -1.39813811e+00 -1.31297673e+00]
 [-1.50652052e+00  1.06445364e-01 -1.28440670e+00 -1.31297673e+00]
 [-1.02184904e+00  1.26346019e+00 -1.34127240e+00 -1.31297673e+00]
 [-5.37177559e-01  1.95766909e+00 -1.17067529e+00 -1.05003079e+00]
 [-1.50652052e+00  8.00654259e-01 -1.34127240e+00 -1.18150376e+00]
 [-1.02184904e+00  8.00654259e-01 -1.28440670e+00 -1.31297673e+00]
 [-1.74885626e+00 -3.56360566e-01 -1.34127240e+00 -1.31297673e+00]
 [-1.14301691e+00  1.06445364e-01 -1.28440670e+00 -1.44444970e+00]
 [-5.37177559e-01  1.49486315e+00 -1.28440670e+00 -1.31297673e+00]
 [-1.26418478e+00  8.00654259e-01 -1.22754100e+00 -1.31297673e+00]
 [-1.26418478e+00 -1.24957601e-01 -1.34127240e+00 -1.44444970e+00]
 [-1.87002413e+00 -1.24957601e-01 -1.51186952e+00 -1.44444970e+00]
```

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
PCA



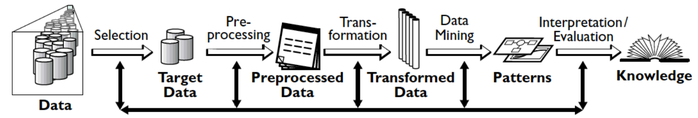
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

```
import pandas
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

#obtenemos la instancia de PCA para dos componentes y obtenemos los
componentes principales
pca = PCA(2)
componentes_principales = pca.fit_transform(x)
conjunto_de_datos_de_PCAs = pandas.DataFrame(data = componentes_principales
            , columns = ['Componente Principal 1', 'Componente Principal 2'])
print(conjunto_de_datos_de_PCAs.tail(4))
```

	Componente Principal 1	Componente Principal 2
146	1.558492	-0.905314
147	1.520845	0.266795
148	1.376391	1.016362
149	0.959299	-0.022284

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
PCA



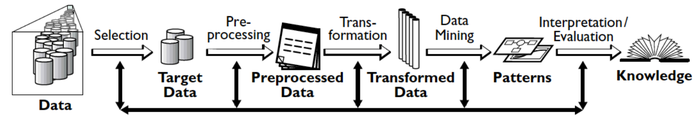
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

```
import pandas
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

#construimos el nuevo conjunto de datos
conjunto_de_datos_final = pandas.concat([conjunto_de_datos_de_PCAs,
conjunto_de_datos[['clase']]], axis = 1)
print(conjunto_de_datos_final.head(10))
```

	Componente Principal 1	Componente Principal 2	clase
0	-2.264542	0.505704	Iris-setosa
1	-2.086426	-0.655405	Iris-setosa
2	-2.367950	-0.318477	Iris-setosa
3	-2.304197	-0.575368	Iris-setosa
4	-2.388777	0.674767	Iris-setosa
5	-2.070537	1.518549	Iris-setosa
6	-2.445711	0.074563	Iris-setosa
7	-2.233842	0.247614	Iris-setosa
8	-2.341958	-1.095146	Iris-setosa
9	-2.188676	-0.448629	Iris-setosa

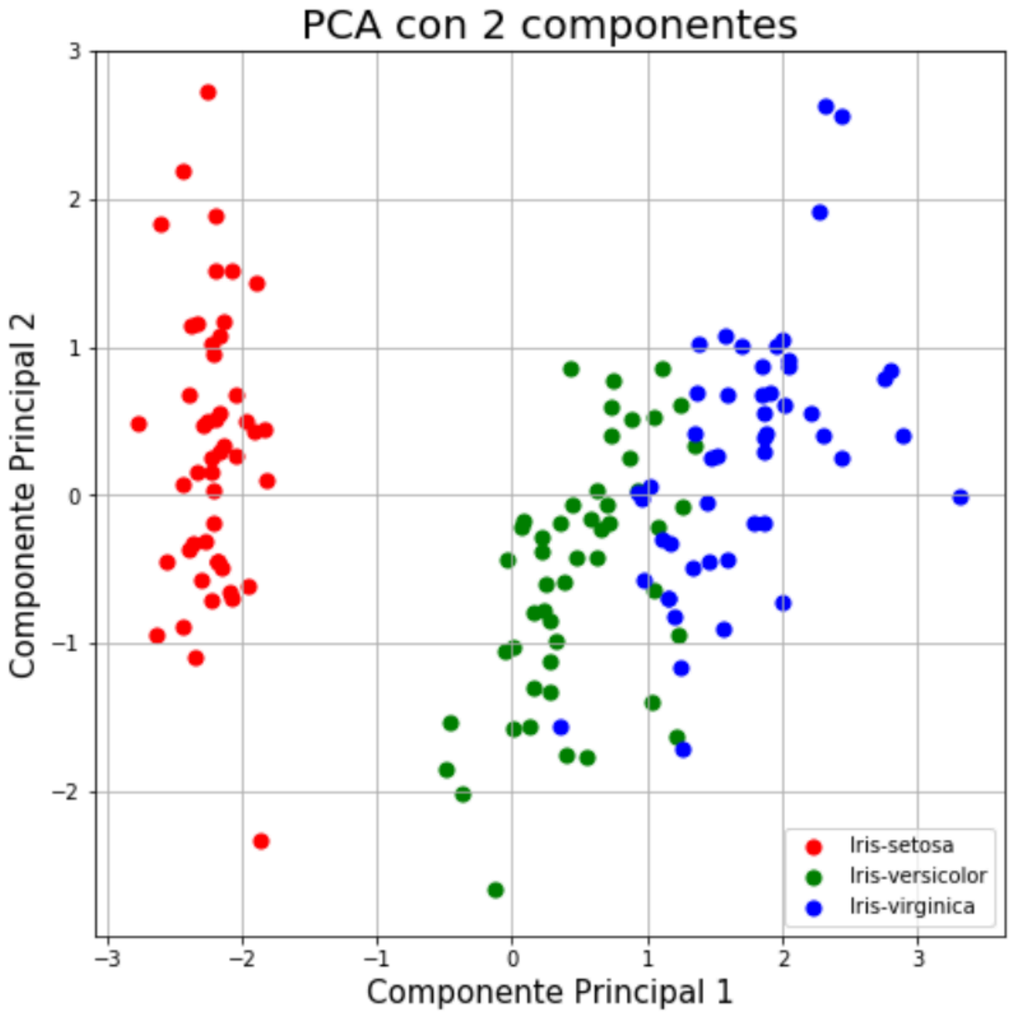
Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
PCA



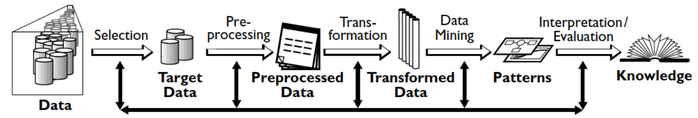
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

```
import pandas
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import matplotlib.pyplot as plot

#proyección en 2D de los componentes principales
fig = plot.figure(figsize = (8,8))
ax = fig.add_subplot(1,1,1)
ax.set_xlabel('Componente Principal 1', fontsize = 15)
ax.set_ylabel('Componente Principal 2', fontsize = 15)
ax.set_title('PCA con 2 componentes', fontsize = 20)
metas = ['Iris-setosa', 'Iris-versicolor', 'Iris-virginica']
colores = ['r', 'g', 'b']
for meta, color in zip(metas, colores):
    indices_de_la_meta = conjunto_de_datos_final['clase'] == meta
    ax.scatter(  conjunto_de_datos_final.loc[indices_de_la_meta, 'Componente
Principal 1']
               , conjunto_de_datos_final.loc[indices_de_la_meta, 'Componente Principal
2']
               , c = color
               , s = 50)
ax.legend(objetivos)
ax.grid()
```



Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
Selección de subconjuntos de atributos

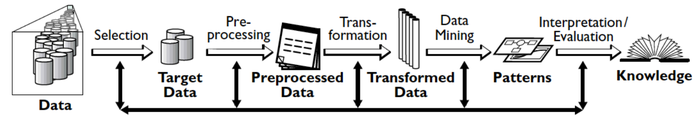


Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- Los conjuntos de datos pueden contener centenares de **atributos**, muchos de los cuales pueden ser **irrelevantes** o **redundantes** para la tarea de aprendizaje.
  - Dejar de lado atributos relevantes o mantener atributos irrelevantes puede causar **confusión** para el algoritmo de **aprendizaje** empleado.
  - Esto puede dar lugar a **modelos** algorítmicos de **mala calidad**.
  - Por otro lado, el volumen agregado de atributos irrelevantes o redundantes puede **ralentizar el** proceso de **aprendizaje**.

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
Selección de subconjuntos de atributos



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

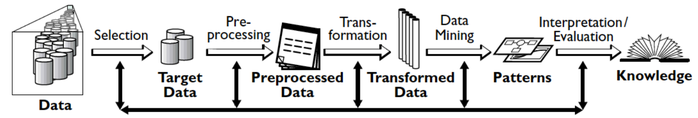
- El objetivo de la selección de subconjuntos de atributos es obtener un **conjunto mínimo de atributos**, de modo que la distribución de probabilidad resultante de las clases sea lo más cercana posible a la distribución original obtenida con todos los atributos.

- La minería en un conjunto reducido de atributos tiene un beneficio adicional:

  - Reduce los atributos que aparecen en los patrones descubiertos, ayudando a hacer los patrones **más entendibles**.

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
Selección de subconjuntos de atributos

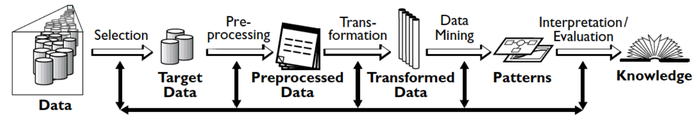


Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- Para *n* atributos, hay 2^n posibles subconjuntos.
  - Una búsqueda exhaustiva del subconjunto óptimo de atributos podría ser muy costosa, especialmente si *n* es un número grande.
  - Por lo tanto, los **métodos heurísticos** que exploran un espacio de búsqueda reducido se suelen usar para la selección de subconjuntos de atributos.
  - Estos métodos suelen ser **"golosos"** en cuanto a que, al buscar en el espacio de atributos, siempre hacen lo que parece ser la mejor opción en ese momento.
  - Su estrategia es hacer una elección **óptima** a nivel **local** con la esperanza de que esto conduzca a una solución global óptima. Son efectivos en la práctica.

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
Selección de subconjuntos de atributos



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- **Selección hacia adelante**
  - El procedimiento comienza con un conjunto vacío de atributos como el conjunto reducido.
  - Se determina el mejor de los atributos originales y se agrega al conjunto reducido.
  - En cada iteración, el mejor de los atributos originales restantes se agrega al conjunto.

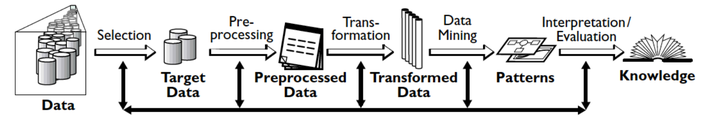
Forward selection

Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

Initial reduced set:
$\{\}$
$\Rightarrow \{A_1\}$
$\Rightarrow \{A_1, A_4\}$
$\Rightarrow$ Reduced attribute set:
$\{A_1, A_4, A_6\}$

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
Selección de subconjuntos de atributos



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- **Eliminación hacia atrás**
  - El procedimiento comienza con el conjunto completo de atributos.
  - En cada paso, elimina el peor atributo que queda en el conjunto.

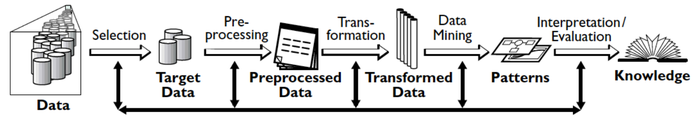
Backward elimination

Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

=> $\{A_1, A_3, A_4, A_5, A_6\}$
=> $\{A_1, A_4, A_5, A_6\}$
=> Reduced attribute set:
$\{A_1, A_4, A_6\}$

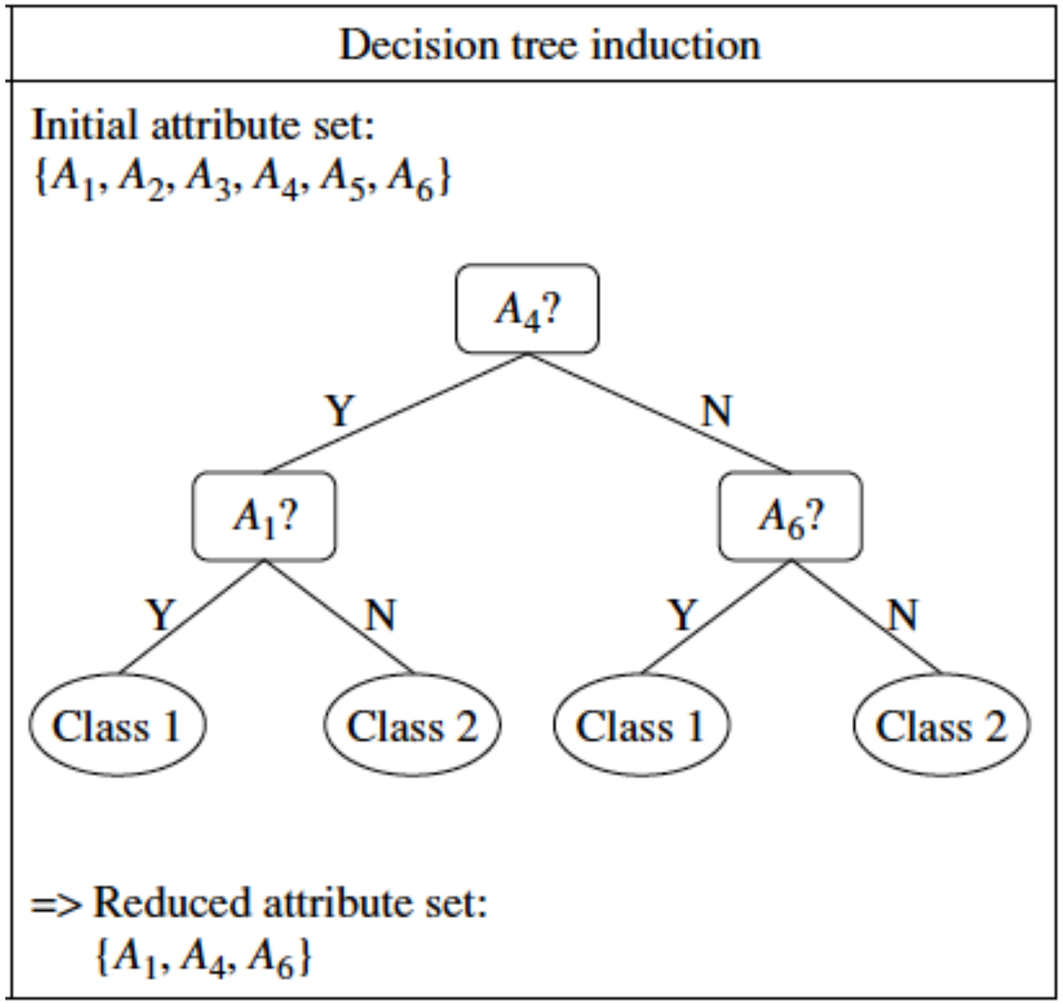
Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
Selección de subconjuntos de atributos



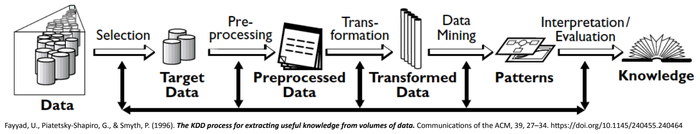
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- **Árbol de decisión**
  - Cuando se utiliza la inducción del árbol de decisión para la selección del subconjunto de atributos, se construye un árbol a partir de los datos dados.
  - Se supone que todos los atributos que no aparecen en el árbol son irrelevantes.
  - El conjunto de atributos que aparecen en el árbol forman el subconjunto reducido de atributos.



Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
Selección de subconjuntos de atributos



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

```
import pandas
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

#cargamos el conjunto de datos
archivo="iris.data"
columnas=['longitud-sépalo', 'ancho-sépalo', 'longitud-pétalo', 'ancho-pétalo', 'clase']
conjunto_de_datos = pandas.read_csv(archivo, names=columnas)

print(conjunto_de_datos.head(5))

#separamos los atributos del atributo meta
x = conjunto_de_datos.iloc[:,0:4].values
y = conjunto_de_datos.iloc[:,4].values

#ejecutamos la selección de atributos
selector = SelectKBest(chi2, k=2)
nuevo_x = selector.fit_transform(x, y)
atributos_seleccionados = selector.get_support(indices=True)

print(nuevo_x)
print(atributos_seleccionados)
```

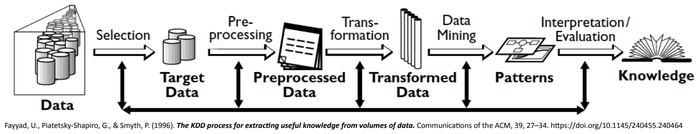
	longitud-sépalo	ancho-sépalo	longitud-pétalo	ancho-pétalo	clase
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

```
[[1.4 0.2]
 [1.4 0.2]
 [1.3 0.2]
 [1.5 0.2]
 [1.4 0.2]
 [1.7 0.4]
 [1.4 0.3]
 [1.5 0.2]
 [1.4 0.2]
```

```
[2 3]
```

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
Selección de subconjuntos de atributos



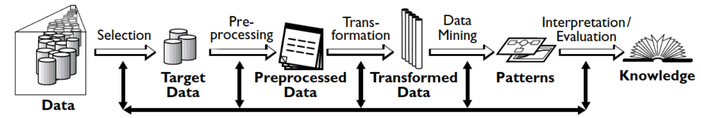
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

```
import pandas
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression

#https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv
#cargamos el conjunto de datos
archivo="diabetes.data"
columnas=['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'clase']
conjunto_de_datos = pandas.read_csv(archivo, names=columnas)
conjunto_de_datos.head(10)

#separamos los atributos del atributo meta
x = conjunto_de_datos.iloc[:,0:8].values
y = conjunto_de_datos.iloc[:,8].values

#selección de los atributos
modelo = LogisticRegression(solver='lbfgs', max_iter=500)
selector = RFE(modelo, 3)
selector.fit(x, y)
atributos_seleccionados = selector.get_support(indices=True)

print(atributos_seleccionados)
```

```
[0 5 6]
```

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos
Reducción de la dimensionalidad
Selección de subconjuntos de atributos



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

```
import pandas
from sklearn.feature_selection import RFE
from sklearn.svm import SVR

#cargamos el conjunto de datos
archivo="diabetes.data"
columnas=['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'clase']
conjunto_de_datos = pandas.read_csv(archivo, names=columnas)
conjunto_de_datos.head(10)

#separamos los atributos del atributo meta
x = conjunto_de_datos.iloc[:,0:8].values
y = conjunto_de_datos.iloc[:,8].values

#selección de los atributos
modelo = SVR(kernel="linear")
selector = RFE(modelo, 3)
selector.fit(x, y)
atributos_seleccionados = selector.get_support(indices=True)

print(atributos_seleccionados)
```

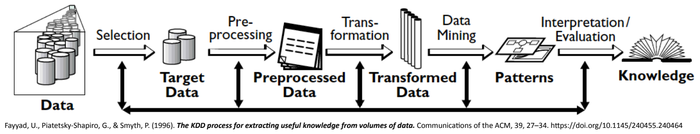
```
[0 5 6]
```

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Reducción de datos

Reducción de la dimensionalidad

Selección de subconjuntos de atributos



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

```
import pandas
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.feature_selection import SelectFromModel

#cargamos el conjunto de datos
archivo="iris.data"
columnas=['longitud-sépalo', 'ancho-sépalo', 'longitud-pétalo', 'ancho-pétalo', 'clase']
conjunto_de_datos = pandas.read_csv(archivo, names=columnas)

print(conjunto_de_datos.head(5))

#separamos los atributos del atributo meta
x = conjunto_de_datos.iloc[:,0:4].values
y = conjunto_de_datos.iloc[:,4].values

#ejecutamos la selección de atributos
selector = ExtraTreesClassifier(n_estimators=50)
selector.fit(x, y)
print(selector.feature_importances_  )

model = SelectFromModel(selector, prefit=True)
nuevo_x = model.transform(x)
print(nuevo_x)
```

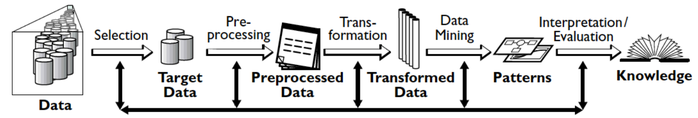
	longitud-sépalo	ancho-sépalo	longitud-pétalo	ancho-pétalo	clase
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

```
[0.10494909 0.07257928 0.42909471 0.39337692]
```

```
[[1.4 0.2]
 [1.4 0.2]
 [1.3 0.2]
 [1.5 0.2]
 [1.4 0.2]
 [1.7 0.4]
 [1.4 0.3]
 [1.5 0.2]
 [1.4 0.2]
```

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Transformación de datos

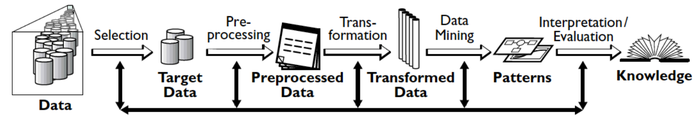


Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- En la transformación de datos, los datos se transforman o se consolidan en formas apropiadas para la minería.
- Las estrategias para la transformación de datos incluyen lo siguiente:
  - **Suavizado:** que sirve para eliminar el ruido de los datos. Las técnicas incluyen suavizado por contenedor, regresión y agrupamiento.
  - **Construcción de atributos** (o construcción de características): donde los nuevos atributos se construyen y agregan a partir del conjunto dado de atributos para ayudar al proceso de aprendizaje.
  - **Agregación:** donde las operaciones de somatización o agregación se aplican a los datos.
  - **Normalización:** donde los datos de los atributos se escalan para que se encuentren dentro de un rango más pequeño, como 1.0 a 1.0, o 0.0 a 1.0.
  - **Discretización:** donde los valores de un atributo numérico (por ejemplo, la edad) se reemplazan por etiquetas de intervalo (por ejemplo, 0-10, 11-20, etc.) o etiquetas conceptuales (por ejemplo, juvenil, adulto, adulto mayor).

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Transformación de datos
Normalización

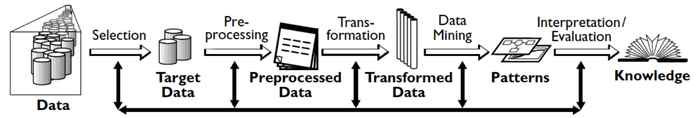


Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- La unidad de medida utilizada puede afectar el análisis de datos.
  - Por ejemplo, cambiar de metros a pulgadas, o de kilogramos a libras, puede llevar a resultados muy diferentes.
- Expresar un atributo en unidades más pequeñas conducirá a un rango más amplio para ese atributo, y por lo tanto tenderá a otorgar mayor efecto o "ponderación" a ese atributo.
- Para ayudar a evitar la dependencia en la elección de las unidades de medida, los datos deben normalizarse o estandarizarse.
  - Esto implica transformar los datos para que se encuentren dentro de un rango más pequeño o común, como [-1, 1] o [0.0, 1.0].

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Transformación de datos
Normalización

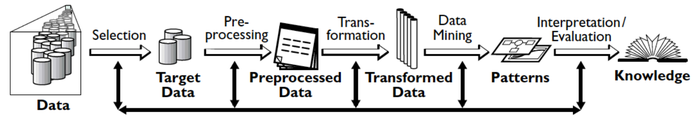


Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- La normalización de los datos intenta dar a todos los atributos un **peso igual**.
- La normalización es particularmente útil para los algoritmos de **clasificación** que involucran redes neuronales o mediciones de distancia, como el vecinos más cercanos y el agrupamiento.
- Para los métodos basados en la **distancia**, la normalización ayuda a evitar que los atributos con rangos inicialmente grandes (por ejemplo, sueldo) superen a los atributos con rangos inicialmente más pequeños (por ejemplo, atributos binarios).

Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Transformación de datos
Normalización

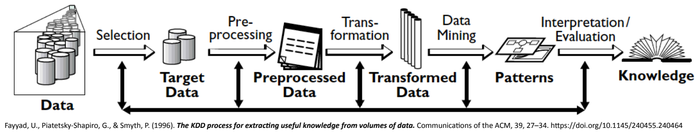


Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). ***The KDD process for extracting useful knowledge from volumes of data.*** Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

- Métodos para normalización de datos:
  - **Normalización min-max:** realiza una transformación lineal sobre los datos originales. $$v_i' = \frac{v_i - min_A}{max_A - min_A}(nuevo\_max_A - nuevo\_min_A) + nuevo\_min_A$$
  - **Normalización z-score:** los valores para un atributo, A, se normalizan según la media y la desviación estándar de A. $$v_i' = \frac{v_i - \overline{A}}{\sigma_A}$$
  - **Escala decimal:** se normaliza moviendo el punto decimal de los valores del atributo A. El número de puntos decimales movidos depende del valor absoluto máximo de A. $$v_i' = \frac{v_i}{10^j}$$

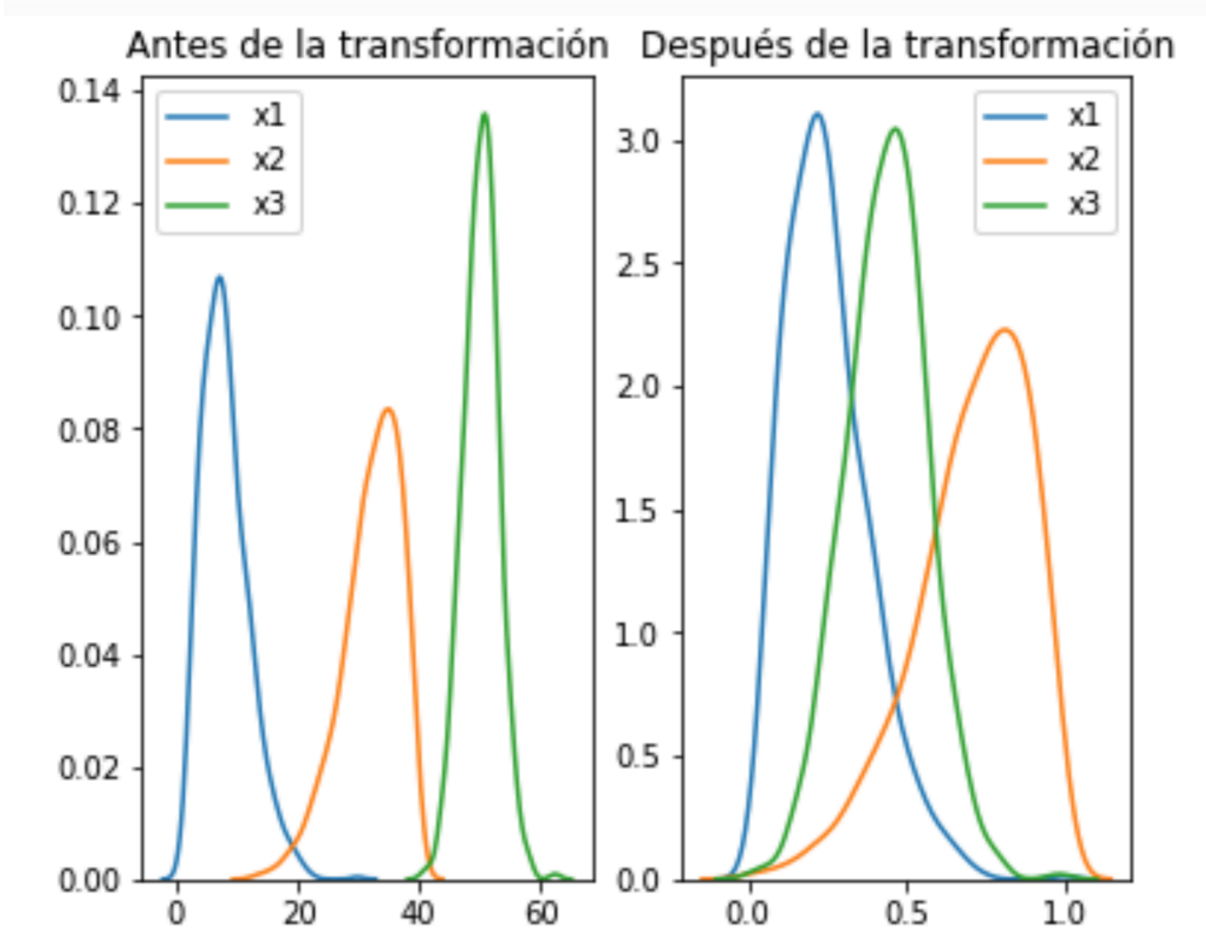
Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Transformación de datos
Normalización



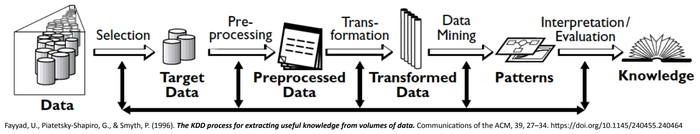
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *The KDD process for extracting useful knowledge from volumes of data.* Communications of the ACM, 39, 27–34. https://doi.org/10.1145/240455.240464

```
import pandas
import numpy
from sklearn import preprocessing
import matplotlib
import matplotlib.pyplot as plot
import seaborn as sns

conjunto_de_datos = pandas.DataFrame({
    #sesgo positivo
    'x1': numpy.random.chisquare(8, 1000),
    #sesgo negativo
    'x2': numpy.random.beta(8, 2, 1000) * 40,
    #sin sesgo
    'x3': numpy.random.normal(50, 3, 1000)
})

#escala = preprocessing.MinMaxScaler(feature_range=(-1, 1))
escala = preprocessing.MinMaxScaler()
conjunto_de_datos_transformado = escala.fit_transform(conjunto_de_datos)
conjunto_de_datos_transformado = pandas.DataFrame(conjunto_de_datos_transformado,
columns=['x1', 'x2', 'x3'])

fig, (ax1, ax2) = plot.subplots(ncols=2, figsize=(6, 5))
ax1.set_title('Antes de la transformación')
sns.kdeplot(conjunto_de_datos['x1'], ax=ax1)
sns.kdeplot(conjunto_de_datos['x2'], ax=ax1)
sns.kdeplot(conjunto_de_datos['x3'], ax=ax1)
ax2.set_title('Después de la transformación')
sns.kdeplot(conjunto_de_datos_transformado['x1'], ax=ax2)
sns.kdeplot(conjunto_de_datos_transformado['x2'], ax=ax2)
sns.kdeplot(conjunto_de_datos_transformado['x3'], ax=ax2)
plot.show()
```

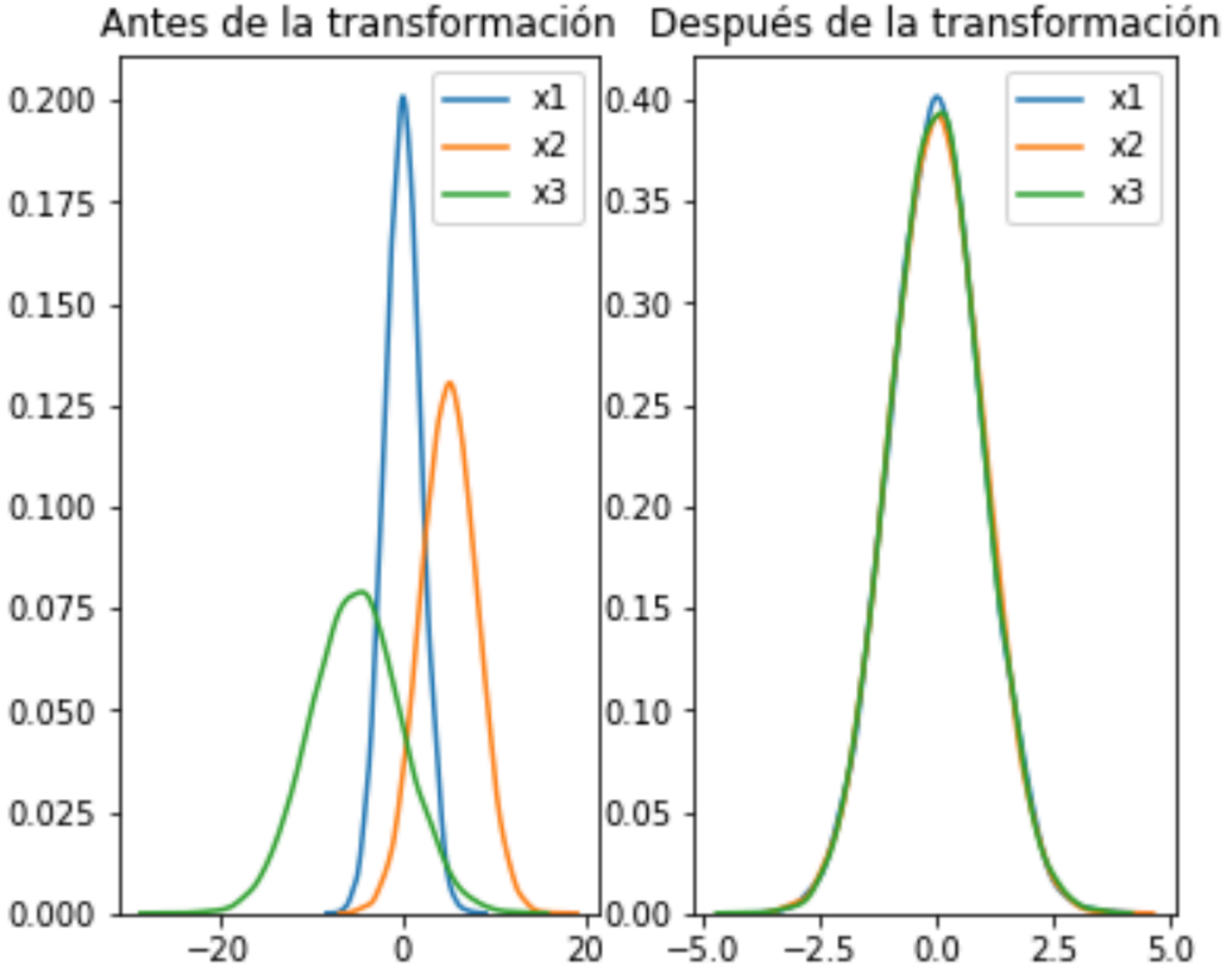


Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Transformación de datos
Normalización




```
import pandas
import numpy
from sklearn import preprocessing
import matplotlib
import matplotlib.pyplot as plot
import seaborn as sns

numpy.random.seed(1)
conjunto_de_datos = pandas.DataFrame({
    'x1': numpy.random.normal(0, 2, 10000),
    'x2': numpy.random.normal(5, 3, 10000),
    'x3': numpy.random.normal(-5, 5, 10000)
})

escala = preprocessing.StandardScaler()
conjunto_de_datos_transformado = escala.fit_transform(conjunto_de_datos)
conjunto_de_datos_transformado = pandas.DataFrame(conjunto_de_datos_transformado,
columns=['x1', 'x2', 'x3'])

fig, (ax1, ax2) = plot.subplots(ncols=2, figsize=(6, 5))
ax1.set_title('Antes de la transformación')
sns.kdeplot(conjunto_de_datos['x1'], ax=ax1)
sns.kdeplot(conjunto_de_datos['x2'], ax=ax1)
sns.kdeplot(conjunto_de_datos['x3'], ax=ax1)
ax2.set_title('Después de la transformación')
sns.kdeplot(conjunto_de_datos_transformado['x1'], ax=ax2)
sns.kdeplot(conjunto_de_datos_transformado['x2'], ax=ax2)
sns.kdeplot(conjunto_de_datos_transformado['x3'], ax=ax2)
plot.show()

print(conjunto_de_datos.head(5))
print(conjunto_de_datos_transformado.head(5))
```



Han, Kamber, Pei, Kamber, Micheline, and Pei, Jian. ***Data Mining Concepts and Techniques.*** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, Mass.: Elsevier, 2012.

Verificación de competencias adquiridas

- Al finalizar la sesión el alumno discutirá las diferentes tareas que se se deben realizar en las fases de **selección**, **pre-procesamiento** y **transformación** del proceso de descubrimiento de conocimiento en base de datos.

- Al finalizar la sesión el alumno ejecutará tareas de **transformación de datos** considerando la **reducción de datos**.

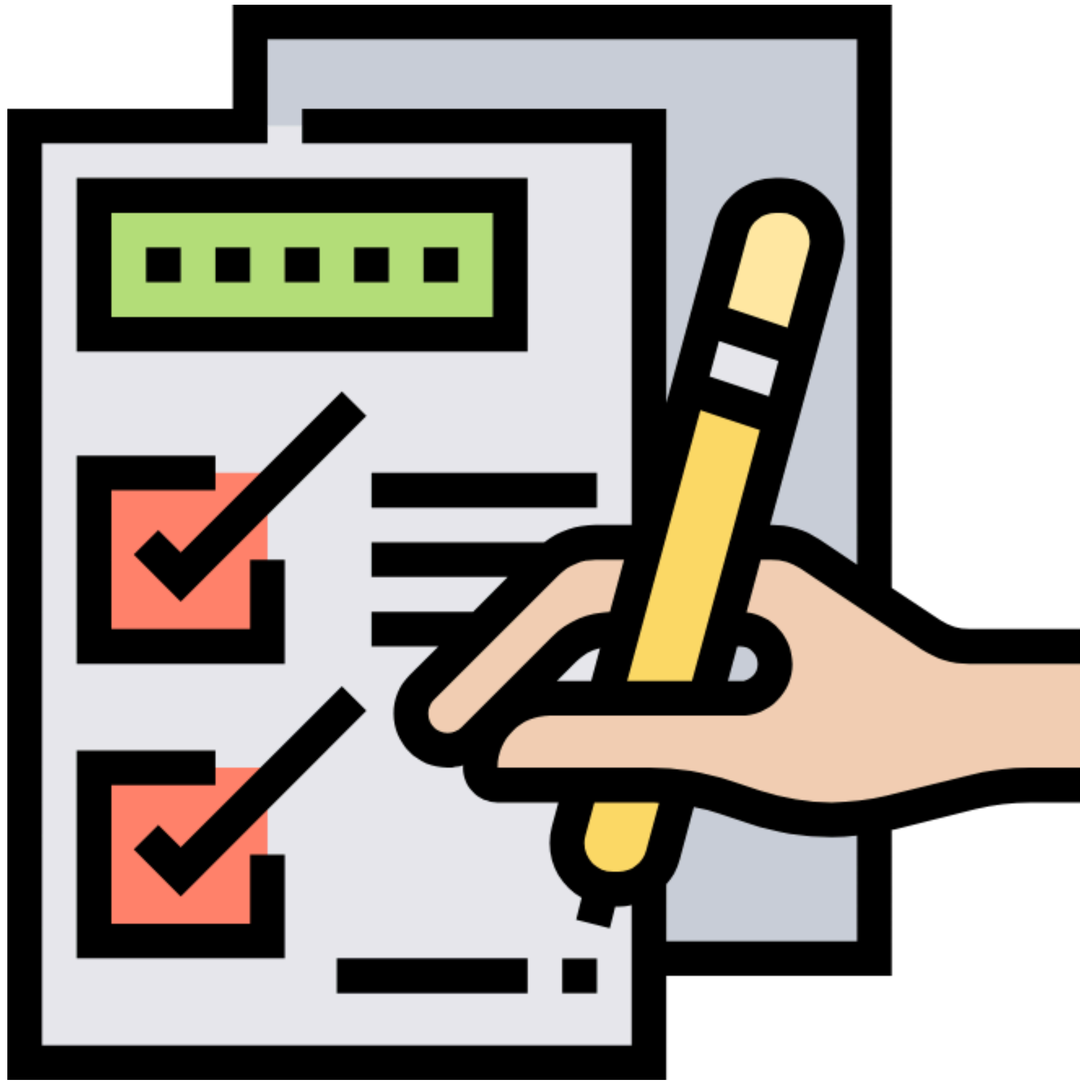


Imagen obtenida a través de la URL https://www.flaticon.com/free-icon/checklist_1467445