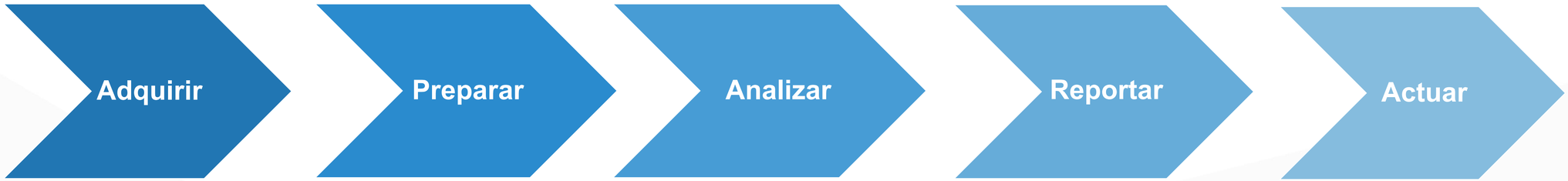
MACHINE LEARNING APLICADO

Exploración de datos

UNIVERSIDAD SIMÓN BOLÍVAR
BARRANQUILLA Y CÚCUTA - COLOMBIA | VIGILADA MINEDUCACIÓN

FACULTAD DE INGENIERÍAS

Proceso del Machine Learning



Proceso del Machine Learning

1. Adquisición

- Identificar fuentes de información
- Recolectar los datos
- Integrar los datos

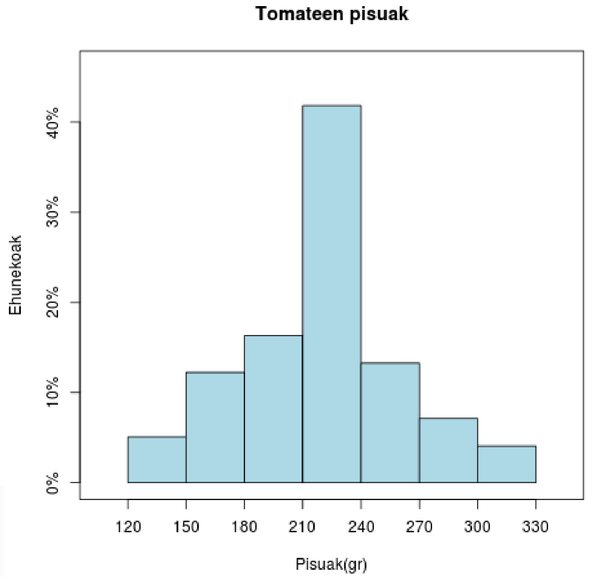
2. Preparación

- Exploración
- Pre-procesamiento

Etapa de Preparación

- Se realizan dos grandes tareas: exploración de los datos y pre-procesamiento

- **Exploración**: se hace para entender los datos:
  - Se busca identificar correlaciones, tendencias y outliers (datos atípicos)
  - Se pueden calcular valores estadísticos importantes como la media, la mediana, la moda, rango y desviación estándar.
  - Se pueden visualizar los datos con gráficos como histogramas, gráficos de dispersión y mapas de calor.



Exploración de datos

❖Los datos que se utilizan en el machine learning tienen la siguiente estructura:

Variables

Muestras

ID	Edad	Peso (Kg)	Estatura (m)
1	20	65	1.55
2	30	70	1.64
3	25	60	1.62
4	40	72	1.70

Exploración de datos

- Variables: también se les conoce como características, campos o atributos. Corresponden a las columnas del dataset.

- Muestras: también se les conoce como registros u observaciones. Corresponden a las filas del dataset.

Tipos de datos

- Numéricos:
  - Ejemplo: 5, 10.5

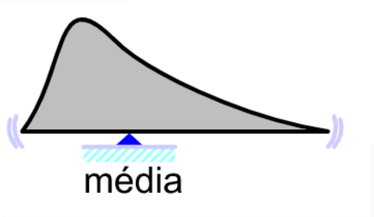
- Categóricos:
  - Ejemplo: Masculino o Femenino.

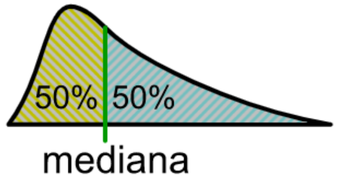
- Otros: fechas, cadenas

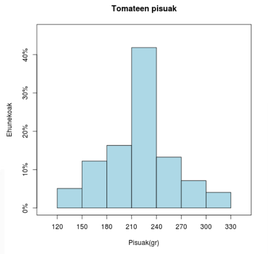
Estrategias de exploración

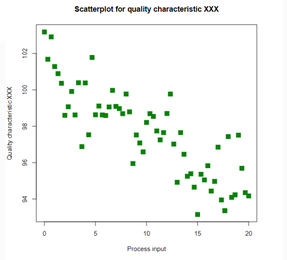
Maneras de explorar los datos

- Estadísticas de Resumen
- Visualización









Estadísticas de resumen

- **Media**: corresponde al promedio de las observaciones. Si X_1, X_2 .... X_n constituyen una muestra de n observaciones, la media aritmética se define de la siguiente manera:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

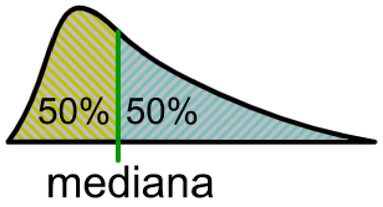
- **Mediana**: es el valor que ocupa la posición central en un conjunto de datos, ordenados en forma creciente o decreciente.

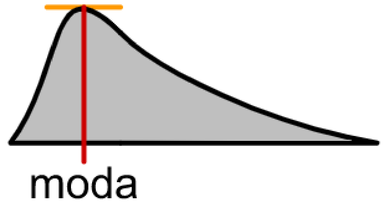
Estadísticas de resumen

- **Mediana:** Para calcular la mediana, primero se deben ordenar los datos. Luego se debe determinar el dato que ocupa la posición (n+1)/2 (cuando n es impar) o la semisuma de los valores numéricos correspondientes a las dos observaciones centrales (cuando n es par).

- **Moda**: Es el valor que ocurre con mas frecuencia.

- **Rango**: Es la diferencia entre las observaciones mayor y menor.



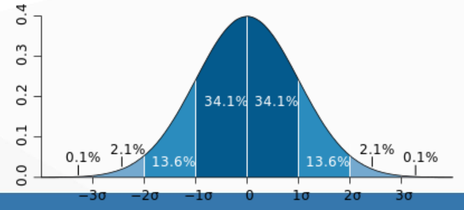


Estadísticas de resumen

- **Varianza:** Es una medida de variabilidad o de dispersion. La varianza muestral se calcula de la siguiente manera:

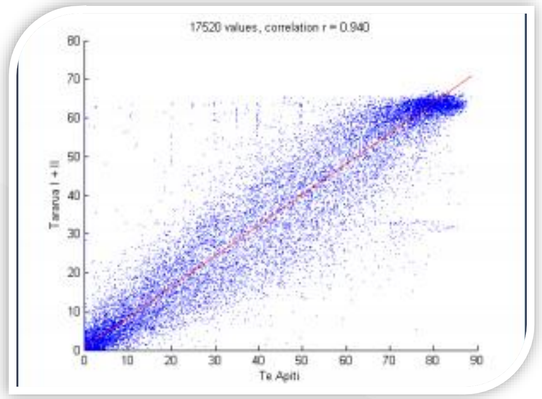
$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

- **Desviación estándar:** Es la raiz cuadrada de la varianza. La varianza y la desviación estándar miden la dispersión "promedio" en torno a la media, es decir cómo fluctúan las observaciones mayores por encima de la media y cómo se distribuyen las observaciones menores por debajo de ella.

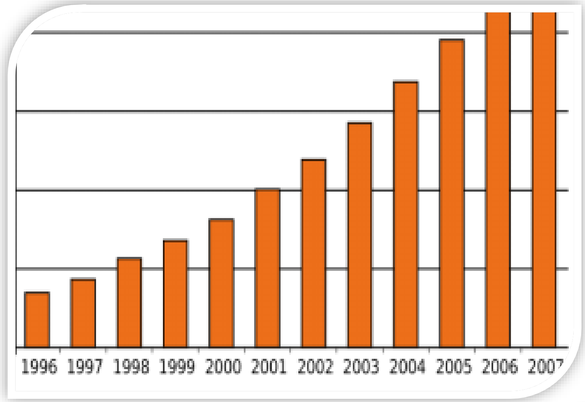


Visualización de datos

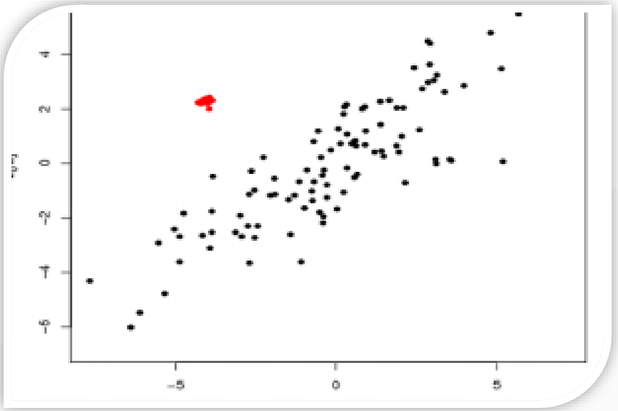
❖ Aspectos a observar:



Correlaciones



Tendencias



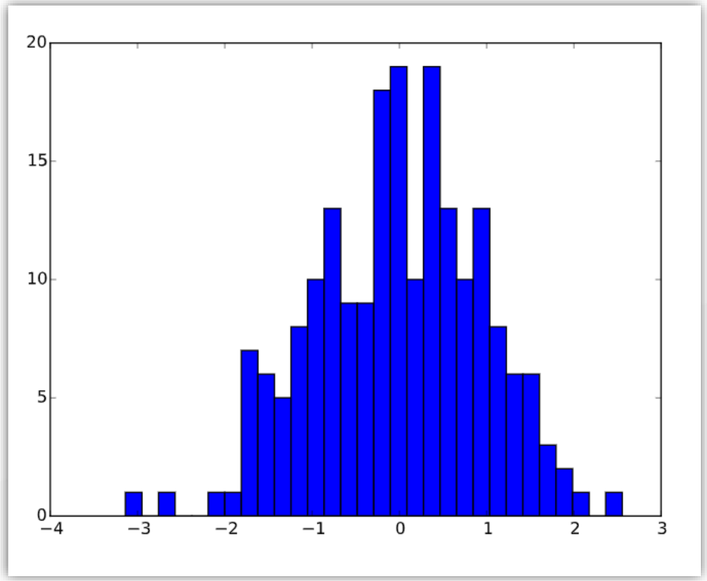
Outliers

Visualización de datos

Histograma

- Un histograma es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados.

- En el eje vertical se representan las frecuencias, y en el eje horizontal los valores de las variables.



Visualización de datos

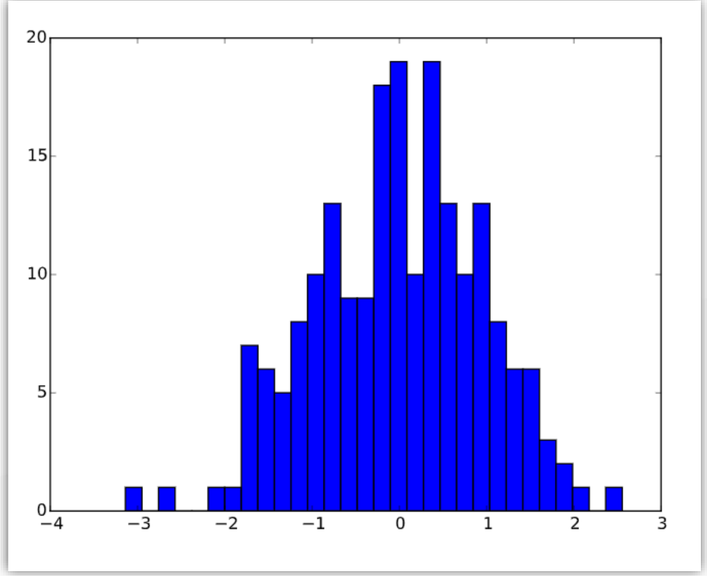
Histograma

- Primero se calcula el rango de los datos y este se divide entre el número de clases $n_{c.}$

- El número de ocurrencia en cada clase es definido como frecuencia. El ancho de cada clase es fijo, para facilitar la interpretación.

- Se recomienda un número de clases entre 5 y 25. Se puede calcular el número ideal con la siguiente formula:

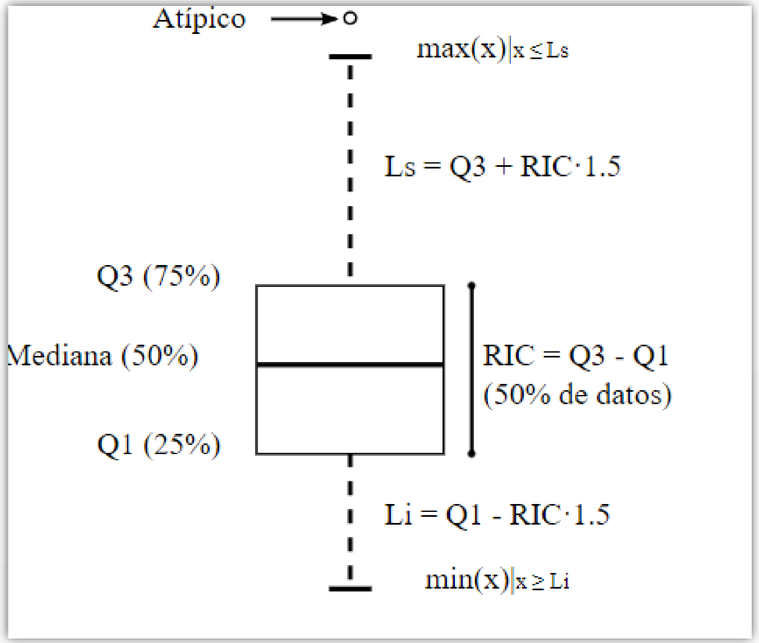
$$n_c = 1 + 3.3\ log10\ n.$$



Visualización de datos

Gráfico de Caja y Bigotes (Box plot)

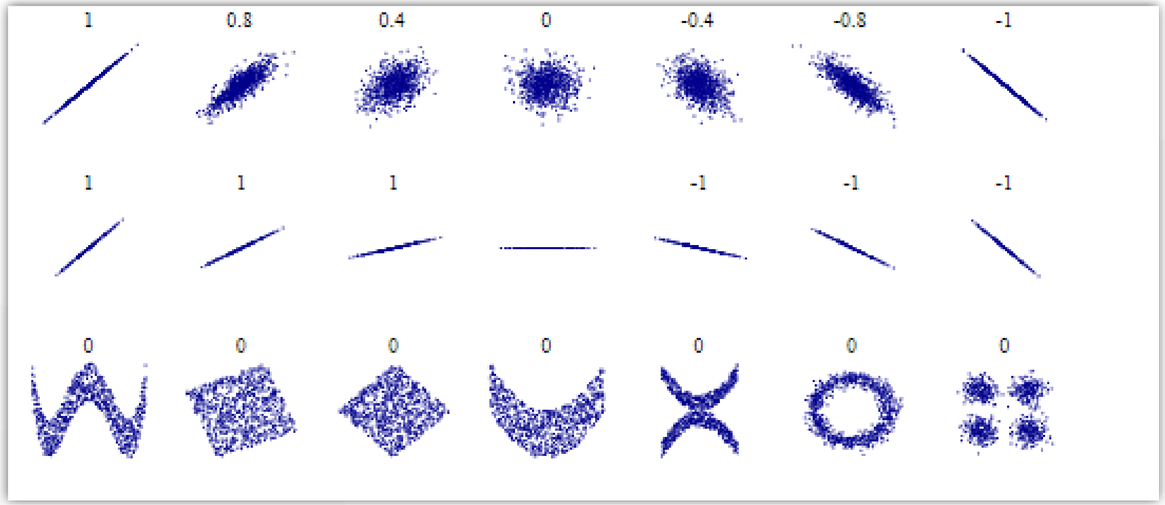
- Los 3 cuartiles principales (Q1, Q2, Q3) son graficados en forma de caja para representar la variación de los datos.
- El Box Plot esta compuesto por 5 cantidades estadísticas fundamentales, Q1, Q2, Q3, máximo y mínimo.
- Este gráfico es útil para identificar outliers o datos atípicos.



Visualización de datos

Gráfico de Dispersión (scatter plot)

- Permite gráficar como es la relación entre dos variables y hacer inferencias sobre su coeficiente de correlación.



Visualización de datos

Covarianza ($S_{x,y}$) y Coeficiente de Correlación ($r_{x,y}$)

$$s_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$r_{X,Y} = \frac{1}{n s_X s_Y}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$-1 \leq r_{X,Y} \leq 1$$

¿Preguntas?

Puedes publicar tu pregunta en el foro de Inquietudes y Novedades

O escribir al correo:
silvia.moreno@unisimon.edu.co

Referencias

- Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media. 2017.

- Hastie, Trevor, Tibshirani, Robert y Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. 2016

- Russell, Stuart J. Inteligencia artificial, un enfoque moderno. Segunda Edición. México, Prentice-Hall Hispanoamericana. 2013